Microbial Biodiversity in the southern Indian Ocean and Southern Ocean

Flavia Flaviani

Thesis presented for the degree of Doctor of Philosophy

Department of Molecular and Cell Biology

University of Cape Town

December 2016

Plagiarism Declaration

"This thesis/dissertation has been submitted to the Turnitin module (or equivalent similarity and originality checking software) and I confirm that my supervisor has seen my report and any concerns revealed by such have been resolved with my supervisor."

Name: Flavia Flaviani

Student number: FLVFLA001

Signed by candidate

Signature: Signature removed

Date: 07 December 2016

Certification by Supervisor

In terms of paragraph GP7 of the regulations for the degree of Doctor of Philosophy at the University of Cape Town, I certify that I approve of the inclusion in this thesis of the material already published, or submitted for publication by the candidate Flavia Flaviani.

Signed by candidate

Signature removed

Signature

Edward P. Rybicki, PhD
Professor in Microbiology
Director of the Biopharming Research Unit
Department of Molecular and Cell Biology, Faculty of Science
& Institute of Infectious Disease and Molecular Medicine
University of Cape Town

Abstract

The multi-phylotype and ecologically important community of microbes in aquatic environments ranges from the numerically dominant viruses to the diverse climate-change regulating phytoplankton. Recent advances in next generation sequencing are starting to reveal the true diversity and biological complexity of this previously invisible component of Earth’s hydrosphere. An increased awareness of this microbiome’s importance has led to the rise of microbial studies with marine environmental samples being collected and sequenced daily around the globe. Despite the rapid advancement in knowledge of marine microbial diversity, technical difficulties have constrained the ability to perform basin wide physical and chemical oceanographic assessments in tandem with microbiological screening with the majority of studies only looking at a single component of the microbial community.

In this study the full microbial diversity, from viruses to protists, was characterised within the southern Indian Ocean and the Southern Ocean from a small volume of seawater collected using the same CTD equipment used by oceanographers. Throughout this study it will be demonstrated how this small volume is sufficient to describe the core microbial taxa in the marine environment. The application of a bespoke bioinformatics pipeline, integrated with sequencing replication, improved the description of the dominant core microbiome whilst removing OTUs present due to PCR and sequencing artefacts thereby improving the accurate description of rare phylotypes. Analyses confirmed the dominance of Cyanobacteria, Alphaproteobacteria and Gammaproteobacteria in the pelagic prokaryotic microbiome, while the Stramenopiles-Alveolata-Rhizaria (SAR) cluster dominates the eukaryotic microbiome. A decrease in the SAR community will be reported for the Southern Ocean with a concomitant increase in the haptophyte community. Whilst the virome confirmed the dominance of tailed phages and giant viruses across all stations, there was a significant variation in the caudoviruses and Nucleocytoplasmic Large DNA viruses (NCLDV) across defined biogeographical boundaries. The described method will allow the characterisation of the microbial biodiversity as well as future integration with oceanographic data with a much reduced sampling effort. The characterisation of the whole microbial community from a single water sample will improve the understanding of microbial interactions and represent a step towards in the inclusion of viruses into biogeochemical models.

Acknowledgments

I would like to acknowledge my supervisor Prof. Ed Rybicki and co-supervisors Dr. Declan Schroeder and Dr. Maya Pfaff. For the guidance, support and the intellectual challenge you gave to me. I am thankful for the opportunity and the trust; it has been a journey that I will always cherish. A special thanks to Prof. Rybicki for being such an inspiration and a great mentor, for showing me my strength and allowing me to believe in myself as well as showing me the joy of communicating science to the world.

For samples collections, statistical advices and substantial support I would like to acknowledge doctors Cecilia Balestreri, Jo Schroeder and Karen Lebret.

I would like to acknowledge the University of Cape Town's ICTS High Performance Computing team for providing the computational infrastructure and special thanks to Andrew Lewis and Timothy Carr for the technical support.

I am grateful to the Linux support provided by Alec Colebrook-Clark during my time at the Marine Biological Association, as well as the technical support provided by Nick Bloomer and Scott Middleton.

For the help and support in the lab I would like to thank Madhu Chauhan, Shakiera Sattar and Tatiana Millard from the University of Cape Town; Matt Hall, Angela Ward and Claire Jasper from the Marine Biological Association of the UK for all your help and support during my time in both laboratories.

This project would have not be possible without the support of the National Research Foundation (NRF) grant to Prof. Ed Rybicki (CPR20110717000020991) and the funding by FP7-OCEAN-2011 call, MicroB3 (grant number 287589) to Dr. Declan Schroeder.

My time spent between UCT and the MBA has granted me the opportunity to meet so many amazing people. I am immensely grateful to Hayley Evers-King and Ben Loveday for the long hours chatting about my project, the oceanography conversations and for showing me how beautiful and special South Africa is, my time there would have not been the same without you.

To my friends Aleyo Chabeda, Hazel Dickens, Megan Coates, Amy Betzelberger, Cecilia Balestreri, Karen Lebret, Jo Schroeder and Beatriz De Francisco thank you for being there for me and listening to me whine while showing me that there is always a way. You were always able to cheer me up when I was down, dragging me out for a chat and some food.

Thanks to Thomas and Rebecca Jackson, Stefan Simis, Marianne Kettunen, Robert Camp, Silvia Pardo, Aser Mata, Emily Mc Gregor and Silvana Mallor Hoya. You guys have been my surrogate family, some from day one some more recently, but you have all provided incredible moral support and chats in the time of difficulties, showing me what friendship is. It's thanks to all of you that I have enjoyed both my time in South Africa and the UK, you filled my life with laughter, adventure, travels and love.

To my family in Sardinia whose support has been constant despite the distance, and for reminding me that no matter where I was, home was always warmer.

Final and enormous thanks to Matt Hall, for your support through these years, the long conversations and the help reading my drafts, this would have not been possible without you. Thank you for reminding me that I am worth it and that someday I will be able to take down that little monster that keeps telling me I am not good enough. You've always been there for me and I am looking forward to an easier time at your side.

Table of Contents

Page

List of Abbreviations and Acronims

ACE - Abundance-based coverage estimator

ACC - Antarctic Circumpolar Current

ADT - Absolute dynamic topography

Anova - Analysis of variance

Avg - Average

APF - Antarctic Polar Front

ARC - Agulhas Return Current

BLAST - Basic Local Alignment Search Tool

blastn - nucleotide BLAST

blastx - translated BLAST (protein databases using a translated nucleotide query)

bp - base pair

°C - degrees Celsius

CTD - Conductivity, temperature, and depth

CO_2 - Carbon dioxide

CO_3 - Carbon trioxide (Carbonate)

Db - database

DNA - Deoxyribonucleic acid

dNTPs – deoxynucleotides

e.g. - exempli gratia

eDNA - Environmental DNA

ENA - European Nucleotide Archive

Etc. - et cetera

HCO_3 - Hydroxidodioxidocarbonate (hydrogen carbonate)

HT1 – Hybridization buffer

i.e. - id est

ICoMM - International Census of Marine Microbes

ICTS - Information and Communication Technology Services

ICTV - International Committee on Taxonomy of Viruses

ITS - International Temperature Scale

Kb - Kilobases

Lat – Latitude

Lon - Longitude

LUCA - Last Universal Common Ancestor

m – meter

Max – maximum

Mb - Megabases

MBA - The Marine Biological Association of the UK

Min - Minimum

$MgCl_2$ - Magnesium chloride

ml - millilitre

mM - millimolar

NASA - National Atmospheric and Space Administration

ng – nanogram

NCBI - National Center for Biotechnology Information

NCLDV - Nucleocytoplasmic Large DNA Virus

NGS - Next Generation Sequencing

NH_4 - Ammonium

nMDS - Non-metric multidimensional scaling

NO_2 - Nitrogen dioxide (Nitrite)

NO_3 - Nitrogen trioxide (Nitrate)

ORF - Open Reading Frame

OTU - Operational Taxonomic Unit

pCO_2 - Partial pressure CO_2

PCR - Polimerase Chain Reaction

PERMANOVA - Permutational multivariate analysis of variance

pmol – picomolar

ppmv - parts per million by volume

PSS - Practical Salinity Scale

P_4 - Tetraphosphorus

RefSeq - NCBI Reference Sequence Database

RNA - Ribosomal ribonucleic acid

rRNA - Ribosomal ribonucleic acid

R/V - Research Vessel

SAR - Stramenopiles, Alveolata, Rhizaria

SBS - Sequencing by synthesis

SDS – Sodium dodecyl sulfate

S. obs - species observed

SYBRG - Syber Green

UCT - University of Cape Town

UK – United Kingdom

u/ µl - units per microliter

µatm - microatmosphere

µl – microliter

µm – micrometer

µmol/l – micromole per liter

µmol/Kg SW - micromole per kilogram of seawater

List of Tables

Page

List of Figures

Page

Chapter 1: Literature Review

1.1 Introduction

Earth is 4.54 (±0.05) billion years old (Dalrymple, 2001) and, whilst there is some debate over the age of life, microfossils have been discovered from 3.5 billion years ago with evidence of biogenic processes from over 4 billion years ago (Bell *et al.*, 2015). As Earth cooled, with the formation of a solid crust and water, chemical processes that favoured the aggregation of molecules and compounds resulted in the development of early non-cellular confined life within the primordial soup, leading towards the genesis of the Last Universal Common Ancestor (LUCA) (Glansdorff *et al.*, 2008; Koonin *et al.*, 2006). Merging between an archaeon and a bacterium brought about the development of eukaryotic cells (Koonin *et al.*, 2006). New hypotheses are proposing a virus-like primordial genetic-system (Koonin *et al.*, 2006); this novel evolutionary scenario probably represents the missing link to the beginnings of cellular life. Nonetheless, numerous debates are ongoing on the inclusion of viruses in the tree of life (Forterre, 2006; Claverie, 2006; Koonin *et al.*, 2006; Forterre, 2010). Indisputably viruses were, and still are, playing a critical role in cellular evolution, with their early effects starting on the cellular lineages derived from LUCA (Bacteria, Archaea and Eukarya) (Forterre, 2010). The origin of life in the oceans can account for the high diversity of microbes that inhabit our planet, providing microbes with billions of years to diversify and evolve, thus enabling colonisation of a multitude of environments across the planet (Margulis and Sagan, 1997).

Microbes, which include viruses, prokaryotes (i.e. Bacteria and Achaea) and small eukaryotes, play important roles in the marine environment and affect all other life on earth. In the late 19th century Louis Pasteur hypothesised that life without microbes would not be possible (Pasteur, 1885). These microscopic organisms that first colonised our planet perform a number of essential roles in the environment, including influencing the carbon and nutrient cycles (Longhurst and Glen Harrison, 1989; Buchan *et al.*, 2014), affecting oxygen production (Pfennig, 1967) and, in the case of viruses and bacteria, are responsible for regulating mortality (Suttle *et al.*, 1990; Proctor and Fuhrman, 1990). These and other processes, in which microbes are involved, have significant roles also as climate regulators (Holligan, 1992). It is therefore essential that we understand the complex ecological interactions between microbes and the environment. Since van Leeuwenhoek's discovery of microbes in 1680 (Smit and Heniger, 1975), many hypotheses have been put forward as to their importance; however proving these has been limited by technology and especially a reliance upon culture-dependent methods for their study. Through advancements in molecular and computational technology we are now gaining better understanding of how this group of organisms evolved, and the key roles they play in biogeochemical cycles. The better understanding of the oceanic systems that microbes inhabit, together with the characterisation of microbial diversity as a whole, will allow the predicting of microbial adaptation and their potential role under different climate change scenarios.

1.2 A water world

"The blue marble" (http://visibleearth.nasa.gov/view.php?id=57723), a photograph of Earth from NASA's 1972 *Apollo 17* mission (Figure 1.1), put into perspective for many the relative percentage of land and water that covers our planet, and the importance of looking after it's ecosystems. Oceans cover over 70% of the Earth's surface (Rahmstorf, 2002) and, with life originating in these oceans, this system is of great importance to our evolution and life.

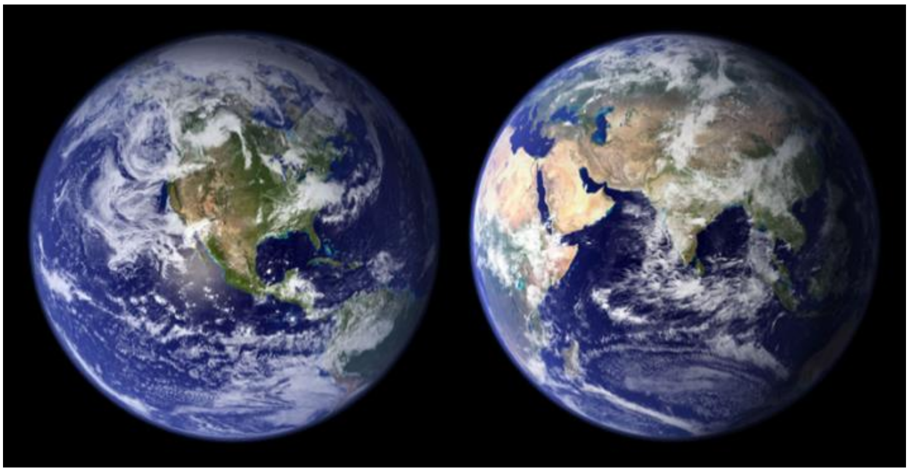


Figure 1.1: The blue marble. Credits: NASA Goddard Space Flight Center Image by Reto Stöckli (land surface, shallow water, clouds). Enhancements by Robert Simmon (ocean colour, compositing, 3D globes, animation). Data and technical support: MODIS Land Group; MODIS Science Data Support Team; MODIS Atmosphere Group; MODIS Ocean Group Additional data: USGS EROS Data Center (topography); USGS Terrestrial Remote Sensing Flagstaff Field Center (Antarctica); Defense Meteorological Satellite Program (city lights).

The oceanic system acts as an essential climate regulator by transporting large amounts of heat, saline water and nutrients via ocean circulation (Houghton, 1996; Macdonald and Wunsch, 1996). Ocean circulation (Figure 1.2) is influenced by a combination of different forces, and therefore an integrated approach is required to understand how they interact. Of these forces wind flows affect predominantly surface waters, whilst fluxes of heat, as cold waters sink, generate and drive the movement of deep-water currents. Changes in salinity created by influxes of fresh water generate the intermediate seawater layer and thermohaline circulation. Finally, gravitational forces, produced by the moon and the sun, regulate mechanical mixing via the tidal cycles (Rahmstorf, 2002).

The Southern Ocean is a high-nutrient and low chlorophyll (HNLC) region, with evidence of iron (Fe) limitation (Popova *et al.*, 2000). Low phytoplankton biomass remains constant throughout the year, and is characterized by several circumpolar quasi-uniform belts that are divided by fronts. Two of these, the Antarctic Polar Front and the Subtropical Front, have been long recognized, while the other fronts were identified during the World Research Programme, which started in 1985 (Ikeda *et al.*, 1989). From North to South they are identified as the Subtropical Front (STF), Sub Antarctic Zone (SAZ), Sub Antarctic Front (SAF), Polar Front (PF) and the Antarctic Zone (AAZ) (Ikeda *et al.*, 1989; Belkin and Gordon, 1996). The Agulhas current is the principal western boundary of the Southern Hemisphere (Lutjeharms and de Ruijter, 1996) and is an important component of the Indian Ocean due to the presence of leakages from this front into adjacent waters. The impact of this system on the global climate was highlighted in a recent review by Beal et al., 2011. Upper warm and salty water from the Indian Ocean enters the South Atlantic Ocean via Agulhas leakages (Donners and Drijfhout, 2004; Beal *et al.*, 2011) regulating the thermohaline circulation cell (Gordon, 1986; Lutjeharms and de Ruijter, 1996) (Figure 1.2). This system represents a key point in global oceanic water circulation because it connects the Atlantic, Indian and Pacific basins (Beal *et al.*, 2011). Presence of a global ocean circulation that transports water around the globe, alongside less distinct marine barriers than terrestrial (Palumbi, 1994, 1992) (i.e. mountain or river), have encouraged the assumption that “everything is everywhere” in the marine environment (Sul *et al.*, 2013; Beijerinck, 1913; Becking, 1934).

At the beginning of the 20th century Martinus W. Beijerinck observed that bacteria appeared ubiquitous and cosmopolitan, and he assumed that therefore they were able to grow everywhere if the conditions were favourable (Beijerinck, 1913). Subsequently in the 1930s the same postulate was refined by Baas Becking who stated that “everything is everywhere, but the environment selects” (Becking, 1934). Since then, an increasing number of studies have demonstrated that marine microbial diversity is structured both by geography and the environment (Williamson *et al.*, 2008; de Vargas *et al.*, 2015; Green and Bohannan, 2006; Feil, 2004; Sul *et al.*, 2013).

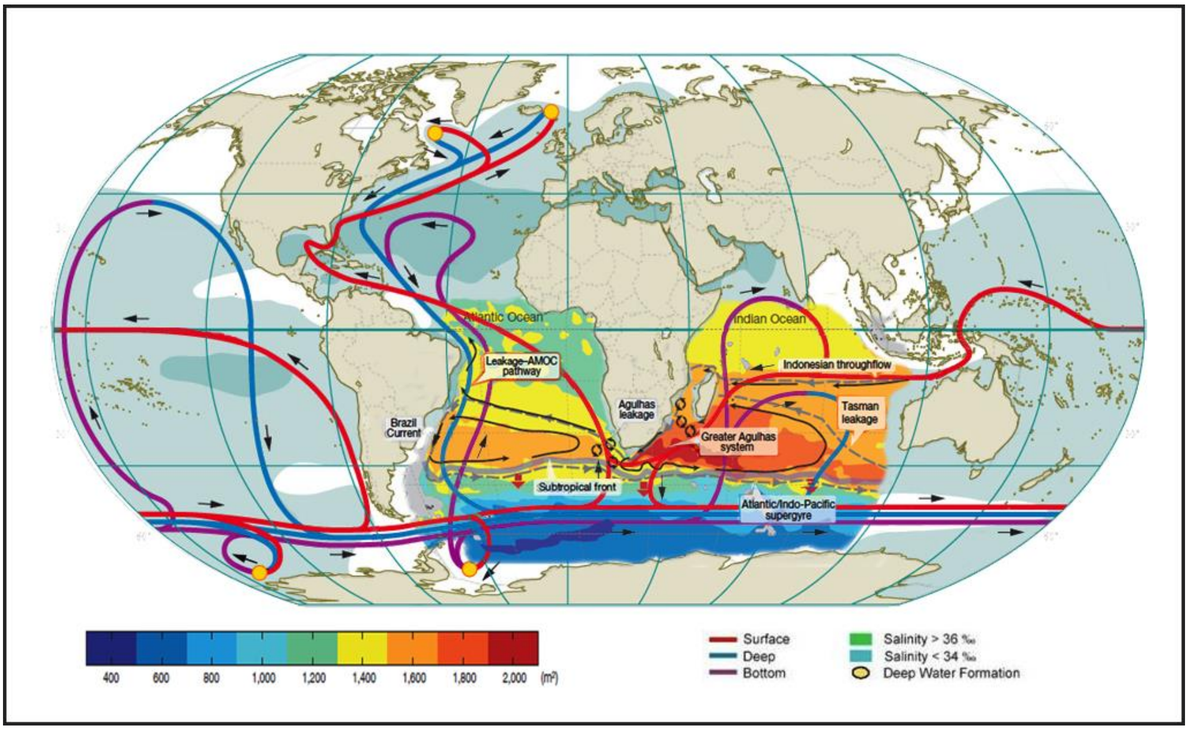


Figure 1.2: Modified from Rahmstorf, 2002 and Beal et al 2011. The "Conveyor belt" simplifies global thermohaline circulation. Shown in red is the surface water, blue the deep water and purple bottom water. Agulhas leakages are shown with black arrows bringing the water from the Indian Ocean through the Agulhas system and then into the Atlantic Ocean.

1.3 The importance of microbes in the oceans

The oceanic biosphere is defined by complex interactions between organisms and their surrounds (Lima-Mendez *et al.*, 2015), with microorganisms playing an important role in its modelling and shaping. Microbes comprise a heterogeneous group of organisms that are grouped together not because of lifestyle, phylogenetic affiliation or similar forms but merely because they are all invisible to the naked eye (Sherr and Sherr, 2000). They constitute more than 90% of the ocean's biomass (Suttle, 2005; Solonenko *et al.*, 2013; Dìez *et al.*, 2001; Fuhrman, 2009), driving almost half of the global primary production (Field, 1998; Cho and Azam, 1990; Azam *et al.*, 1983) and are therefore of great importance for global ecosystems. The microbial community is shaped by the highly variable physical and chemical conditions of the oceanic system, as well as by the presence of predators (Margalef, 1969; Tilman, 1977;

Pedrós-Alió, 2006). In return, they regulate the environment: working as biological engineers of life (Falkowski *et al.*, 2008) shaping the biogeochemical pathways that are critical for the global ocean carbon sequestration and modulating atmospheric CO_2 (Follows and Dutkiewicz, 2011; Follows *et al.*, 2007).

The biological pump is one the ways they shape the marine environment. This mechanism, entirely driven by marine microbes, structures the distribution of fixed carbon, dissolved oxygen and nutrients as well as balancing key factors of the global climate, in a process that removes carbon from the atmosphere and transports it into the deep ocean and seafloor (Pfaff *et al.*, 2014; Longhurst and Glen Harrison, 1989; Ducklow *et al.*, 2001) (Figure 1.3). Photosynthetic organisms present in surface waters capture energy from light, transforming inorganic matter such as CO_2 into organic matter, which is at the base of marine food webs (Buchan *et al.*, 2014). A significant fraction of the newly produced organic matter in the form of particulate organic carbon (POC) is directly used for respiration, and transformed back into CO_2 at the surface and released back in the atmosphere (Herndl and Reinthaler, 2013). Bacteria transform particulate organic matter (POM) into dissolved organic matter (DOM), a nutrient form readily used by other organisms (Buchan *et al.*, 2014; Herndl and Reinthaler, 2013; Ducklow *et al.*, 2001), via the microbial loop. Microbial communities are able to re-use this last form of dissolved organic carbon (DOC as part of the DOM), and consequently increase the consumption of oxygen whilst decreasing the transfer of carbon to higher trophic levels.

Mortality within the microbial community has a very significant contribution from cell death caused by viruses (Wilhelm and Suttle, 1999; Breitbart *et al.*, 2007; Suttle, 2005). Dead cells and debris created from cell lysis, termed ‘marine snow’ (Armstrong *et al.*, 2001; Reinthaler *et al.*, 2009), are responsible for the transport of organic matter into the deepest part of the oceans and the seafloor. Through the viral shunt, in which fixed carbon is shifted ("shunted") to DOM by viral cell lysis, viruses influence the biological pump whereby nutrients and elements sink from surface waters into the thermocline and deep water (Azam *et al.*, 1983; Wilhelm and Suttle, 1999; Suttle, 2007). Marine ecosystems are affected by the increased residence time of carbon and mineral nutrients in the euphotic zone (Moore *et al.*, 2013). In all these processes microbes reduce this time, favouring the regeneration of nutrients for higher trophic levels. As shown in Figure 1.3, marine microbes including viruses, prokaryotes and eukaryotes are widely interconnected playing important roles in the environment and are therefore able to affect, not only the oceanic systems, but also all life on Earth.

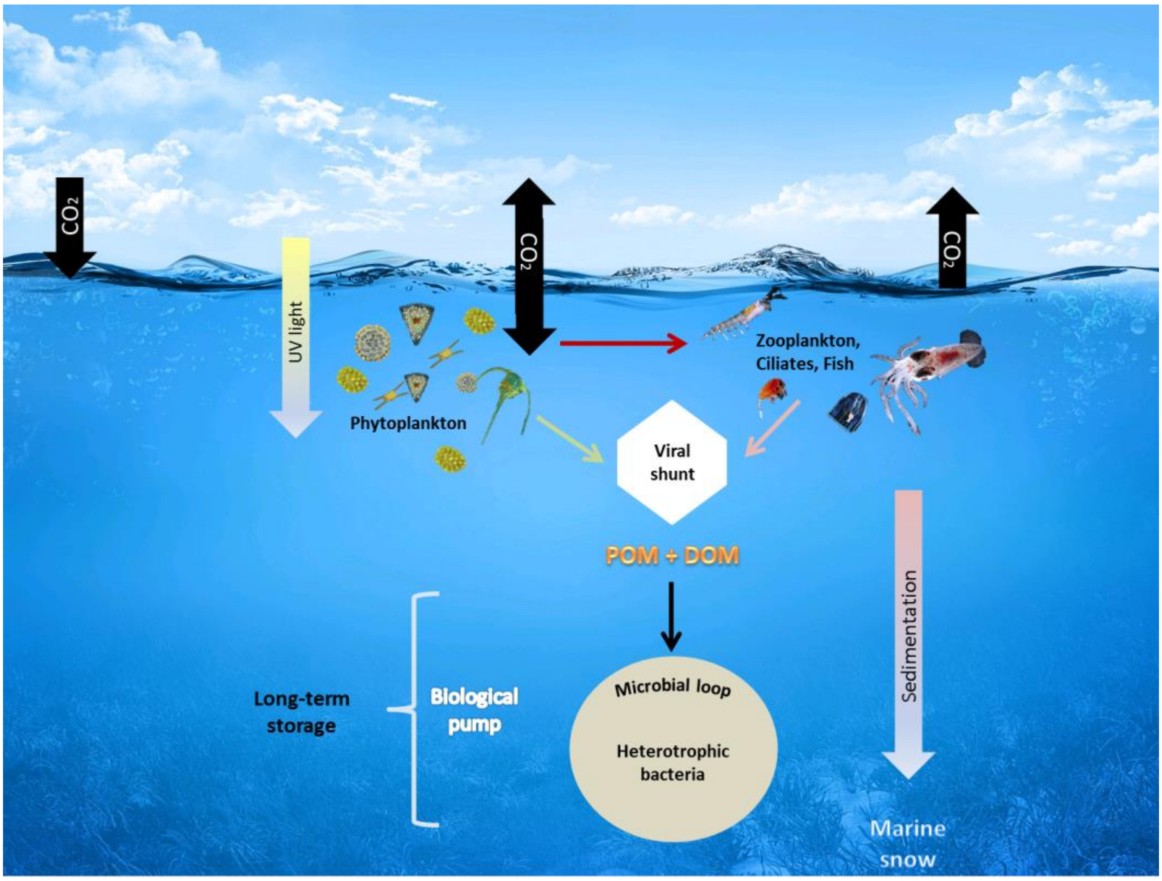


Figure 1.3: Complexity of the roles of microbes in the oceans. From CO_2 sequestration and its use to create organic matter, to the production of oxygen. Microbes, which include viruses, prokaryotes and eukaryotes, are interlinked in the oceans to regenerate nutrients and favour life.

1.4 The missing link

It has been estimated that in a litre of seawater there are 10^9-10^{11} virus particles (Wilhelm and Matteson, 2008), 10^8 prokaryotic (Brown *et al.*, 2009) and 10^6 eukaryotic cells (Whitman *et al.*, 1998), all working together to sustain major biogeochemical processes (see Figure 1.3). Despite microbes global environmental importance, the complex interactions and ecological significance of the relationships within and between biomes are largely unknown. The lack of understanding of viral interactions is independent of the type of environment sampled whether it is marine (Sogin *et al.*, 2006; Brum *et al.*, 2013b), soil (Roesch *et al.*, 2007) or human gut (Turnbaugh *et al.*, 2009).

The majority of studies only investigated a single group within the microbial world, with only 11.2% monitoring two microbial groups simultaneously and 2.2% looking at the interactions between prokaryotes, eukaryotes and viruses (Zinger *et al.*, 2012). For this reason, many of the recent oceanic expeditions were designed in order to collect data about different trophic levels and ecosystem components in a more comprehensive way, attempting to bring to light the complex ecosystem dynamics. Describing and studying the hosts, prokaryotes and eukaryote assemblages, alongside their viruses can help improve our understanding on the roles of the microbiome in a more holistic way.

Over the past 15 years the world's oceans ecosystems have been explored (Figure 1.4) with an increased focus on microbial communities. Expeditions such as the Global Ocean Sampling (2003-2010, http://www.jcvi.org/cms/research/projects/gos/overview/) (Rusch *et al.*, 2007), Tara Ocean Expedition (2009-2012, http://oceans.taraexpeditions.org/) (Sunagawa *et al.*, 2015), Malaspina (2010, http://www.expedicionmalaspina.es/) (Laursen, 2011) and various census programs such as the Earth Microbiome program (Gilbert *et al.*, 2011), and the Micro B3 led Ocean Sampling Day (Kopf *et al.*, 2015) are contributing to the unveiling of marine microbes - but not without limitations. The gaining of knowledge of marine microbes has been slowed down in the past by the fact that the majority of microorganisms cannot be grown under laboratory conditions (Handelsman, 2004) and the information from these laboratory cultures is extremely limited (Follows *et al.*, 2007). However, due to major efforts on sampling the marine environment, together with the advancement of sequencing chemistry and technologies, we are now able to study the marine microbial community without the need of cultivation steps (Loman *et al.*, 2012a).

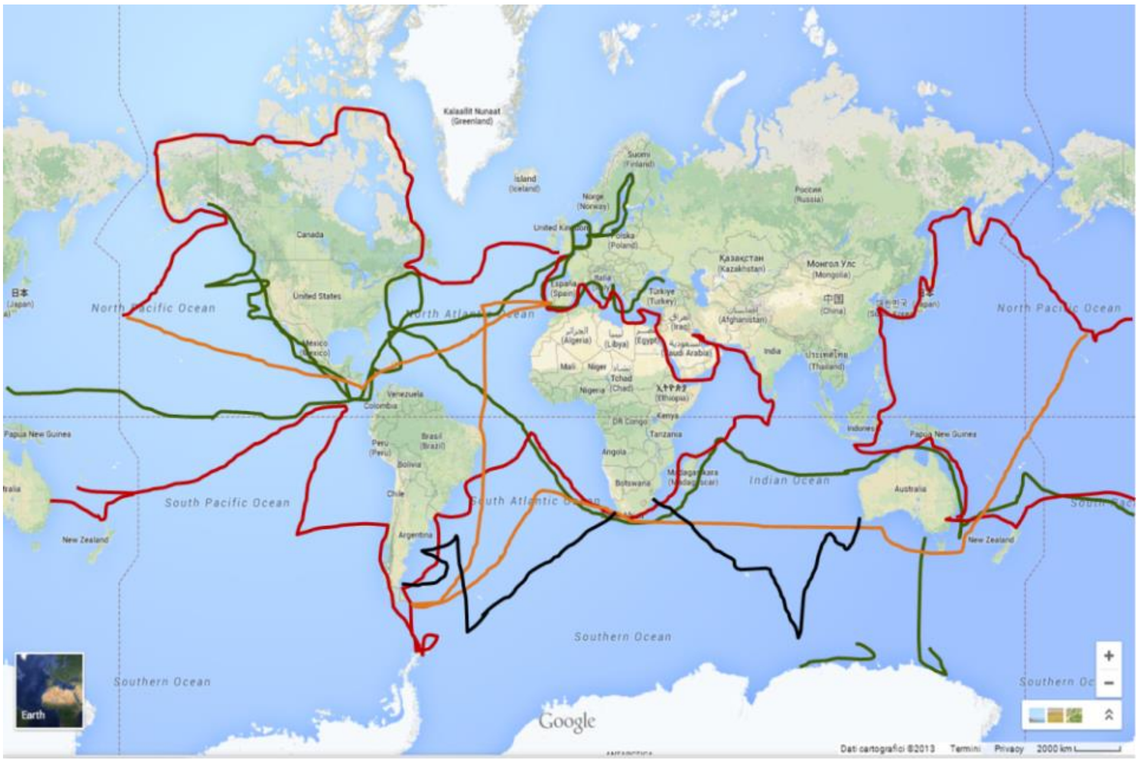


Figure 1.4: Global expeditions tracks. Green: Global Ocean Sampling (2003-2010); red: Tara's Ocean Expedition (2009-2012); orange: Malaspina (2010); black: Great Southern Coccolithophore Belt expedition (2011-2012).

1.5 Marine prokaryotes

Microbes are characterised by greater phylogenetic and physiological diversity than animals or plants and their interactions with the environment are more complex (Pace, 1997). Prokaryotes include two domains: these are Bacteria and Archaea (Woese and Fox, 1977). In the past, the diversity of marine microbes, and specifically prokaryotes, has been calculated through cell counts, which has led to biases in the estimation of microbial abundance in the oceans (Amaral-Zettler *et al.*, 2010). It was only with the use of the 16S rRNA gene in the 1980s (Pace *et al.*, 1986) that it was realised that cultivation was not enough for microbial characterisation. Direct counts from plates had estimated the presence of 100 cells (Amaral-Zettler *et al.*, 2010) for each millilitre of seawater, whilst fluorescent techniques showed an average of 1,000,000 cells per millilitre (Whitman *et al.*, 1998), five orders of magnitude more than estimates through plate counts. This became known as the "great plate-count anomaly" (Staley and Konopka, 1985) which is reinforced further if we take into consideration sequences deposited in global databases. Prior to 2010, more than 10,000 bacterial and archaeal sequences from cultivation based studies, were deposited in databases (Amaral-Zettler *et al.*, 2010). In contrast, culture independent 16S rRNA based studies identified this number to be 100 times higher (Pace, 1997), highlighting the downside of depending on cultivation techniques for estimations of microbial diversity.

Traditional phenotypic characterisation of the prokaryotes (Bergey *et al.*, 1984) has thus been replaced with identification through 16S rRNA gene (Boone *et al.*, 2001). This classification utilises the Phylum as its highest rank which includes: Proteobacteria, Bacteroidetes, Chlorobi, Cyanobacteria, Actinobacteria, Acidobacteria, Firmicutes, Planctomycetes, Verrucomicrobia, Aquificae, Chlamydia, Deferribacteres, Spirochaetes, Fibrobacteres, Nitrospira, Fusibacteria, Chloroflexi, Deinococcus-Thermus, Dictyoglomi and Thermotogae (Ludwig and Klenk). Archaea are formally classified in Crenarchaeota and

Euryarchaeota based on 16S rRNA gene (Bergey *et al.*, 1984), with new studies suggesting the non-monophyly of the Euryarchaeota group (Wolf *et al.*, 2001).

Global marine prokaryotic diversity has a high abundance of Alphaproteobacteria in both surface waters (SRF) and at the deep chlorophyll maximum (DCM) (Sunagawa *et al.*, 2015; Giovannoni *et al.*, 1990; Amaral-Zettler *et al.*, 2010; Zinger *et al.*, 2011). The International Census of Marine Microbes (ICoMM) identified Gammaproteobacteria as the second most abundant group for the aquatic realm (including coastal waters, seamounts, polar waters and open ocean) as well as the pelagic (Amaral-Zettler *et al.*, 2010; Zinger *et al.*, 2011). During the Tara Ocean expedition the second most abundant group identified was Cyanobacteria and Gammaproteobacteria at varying proportions depending on locations. An exception to these results was the south-west Indian Ocean which was dominated by Cyanobacteria taxa, then Gammaproteobacteria and finally Alphaproteobacteria (Sunagawa *et al.*, 2015). The global distribution of Gammaproteobacteria across a variety of marine habitats can be explained by their large phenotypic and phylogenetic diversity (Williams *et al.*, 2010).

1.6 Marine microbial eukaryotes

Marine microbial eukaryotes can be subdivided into three categories based on size: these are picoplankton, which at first included only prokaryotes (0.2-2 µm), nanoplankton (2 - 20 µm) and microplankton (20 - 200 µm) (Sieburth *et al.*, 1978). Cell counts range between 10^3 and 10^5 cells per millilitre of seawater depending on the oligotrophy of the environment (Li, 2009; Sanders *et al.*, 2000), with counts increasing with depth in the water column until the deep chlorophyll maximum is reached, and showing an abrupt decrease below this (Massana, 2011). Similarly to the prokaryotes, sequencing technologies and the advance of molecular techniques helped characterise the community especially the smallest fraction

(Massana and Pedrós-Alió, 2008). Sequencing of environmental genes such as the 18S rRNA are utilised to quantify diversity in this group (Massana, 2011). Furthermore, the use of high throughput sequencing studies with no cloning step (Cheung *et al.*, 2010) are simplifying the process and advancing our understanding.

Fundamentally eukaryotes can be clustered into supergroups (Massana and Pedrós-Alió, 2008) characterised by distinct lineages, mainly protists, with similar phylogenetic characteristics and structure (Adl *et al.*, 2005; Baldauf, 2003; Simpson and Roger, 2004). The supergroup Alveolata is composed of primary producers (Guillou *et al.*, 2008) important in the oceans and dominates marine eukaryotic surveys (Massana, 2011; de Vargas *et al.*, 2015; Amaral-Zettler *et al.*, 2010); it includes four classes: Dinoflagellata, Apicomplexa, Ciliophora and Perkinsea (Guillou *et al.*, 2008). In this group are included novel lineages such as marine alveolates (MALV); recently MALV-I and MALV-II have been reclassified as Syndiniales groups I and II (Horiguchi, 2015). A large number of alveolate species are parasites with the class Apicomplexa characterised exclusively by obligate parasites, whilst ciliates and dinoflagellates can behave as active predators (Guillou *et al.*, 2008). Furthermore, dinoflagellates are known in the oceans for their photosynthetic role (Lessard and Swift, 1986) as well as being responsible of toxic algal blooms (Smayda, 1997; Eberlein *et al.*, 2016).

During the ICoMM survey it was shown that Alveolata, specifically dinoflagellates, dominate across the various water sources analysed (Amaral-Zettler *et al.*, 2010). However, the high frequency of this group has been associated with a bias due to high copy number of the rRNA genes (Zhu *et al.*, 2005). Samples collected during the Tara Ocean expedition for the eukaryotic fraction (de Vargas *et al.*, 2015) showed that the pico-nanoplankton was dominated by photosynthetic dinoflagellates (family Dinophyceae). However, heterotrophic protists showed the highest richness and abundance across all the other size fractions. Parasites of the order Alveolata, known to routinely infect the Dinophyceae, were mainly constituted of the order Syndiniales, specifically the MALV- I and MALV-II clusters (up to 88% of abundance across some stations).

1.7 Marine viruses

The main body of marine viral research began in 1970's, and by the 1990's the potential significance of marine viruses was reported, with hypotheses made as to their function (Bergh *et al.*, 1989; Wilhelm and Suttle, 1999; Culley, 2011). Their role in global biogeochemical cycles is now well established (Fuhrman, 1999; Wilhelm and Suttle, 1999; Suttle, 2005, 2007; Rohwer and Thurber, 2009) as is their impact on the ecological community structure through infection, involvement in host mortality (Wilhelm and Suttle, 1999; Suttle, 2005, 2007) and effect on the transfer of genetic material (Sano *et al.*, 2004; Suttle, 2005). In the last decade the importance of viruses in the marine environment has become clearer and consequently the need to understand their role in this system has grown. Advances in sequencing technology and molecular biology have facilitated the rapid progress in the understanding of the role viruses play in the oceans (Edwards and Rohwer, 2005), but much remains to be discovered.

Viruses are numerically the most abundant biological entities on the planet, with estimates ranging from 10^7 to 10^9 per millilitre of seawater (Martínez Martínez *et al.*, 2014; Williamson *et al.*, 2012, 2008; Bergh *et al.*, 1989). It has been predicted that bacteriophages outnumber their bacterial hosts in the marine environment by an order of magnitude (Bergh *et al.*, 1989; Wommack and Colwell, 2000; Weinbauer, 2004; Wigington *et al.*, 2016). Despite the increasing awareness of the importance of viruses in key biological processes, major bottlenecks on viral diversity and viral roles in marine ecosystems still remain (Roux *et al.*, 2015). The majority (up to 95%) of gene/protein sequences in marine viromes cannot be assigned to known virus genes/proteins (Mizuno *et al.*, 2013; Brum *et al.*, 2013b; Angly *et al.*, 2006; Williamson *et al.*, 2012), causing difficulties in positively identifying viruses within the environment.

In the marine environment the genome size of bacteriophage ranges between 20 and ~250kb (Sandaa, 2008; Steward *et al.*, 2000; Lavigne *et al.*, 2009), while for giant viruses infecting eukaryotes the genome sizes range from 100kb to 2.5Mb (Colson *et al.*, 2013; Yutin and Koonin, 2013; Claverie *et al.*, 2006; Philippe *et al.*, 2013; Iyer *et al.*, 2001). Viruses belonging to the order *Caudovirales* infect bacteria (Ackermann, 2003), and comprise three families: these are viruses in families *Myoviridae* (contractile tails), *Siphoviridae* (non-contractile tails) and *Podoviridae* (short tails) (Ackermann, 2003). Giant viruses that infect marine protists (Blanc-Mathieu and Ogata, 2016) include the Large Nucleocytoplasmic DNA Viruses (NCLDVs), recently proposed to be grouped into the suggested order *Megavirales* (Colson *et al.*, 2013).

1.8 Sequencing technologies

Only 0.1-1% of microbes in the environment have been cultured (Rappé and Giovannoni, 2003; Edwards and Rohwer, 2005), which has limited the number of microbes, and consequently viruses, that can be detected through cultivation techniques such as plaque assays. If on one side the presence of universally of conserved genes such as the 16S and 18S rDNA have facilitated the early exploration of marine microbes without cultivation steps, on the other side the absence of conserved genes in viruses have rendered the study of this group significantly more challenging. The study of microbial communities, including marine microbes, has been limited not only by available technologies but also by the lack of reference genomes (Scholz *et al.*, 2012) as a consequence of difficulties in preparation of laboratory culture for the majority of microbes (Handelsman, 2004). It was only in the 1980s with the utilisation of rDNA sequences (Pace *et al.*, 1986) that the high diversity of marine microbes started to really be discovered. With the advent of high-throughput sequencing (HTS) technologies, such as 454-pyrosequencing and Illumina (Logares *et al.*, 2012), the true diversity of the microbial world was ready to be investigated. It has been demonstrated that both platforms provide comparable representations of the microbial community (Luo *et al.*, 2012; Solonenko *et al.*, 2013) with the two platforms producing analogous results with similarity of ~90% on both the assembled contigs and the unassembled reads (Luo *et al.*, 2012). In Table 1.1 some of the differences between the two technologies are shown. Specifically 454-pyrosequencing was the first next generation sequencing (NGS) platform (Margulies *et al.*, 2005): it generates longer sequence reads whilst Illumina technology produces shorter sequence reads but it offers a better assembly of sequence reads (Luo *et al.*, 2012). Furthermore Illumina offers the broadest utility and lowest cost per read and Mb (Table 1.1) (Glenn, 2011; Liu *et al.*, 2011; Luo *et al.*, 2012). Before the advent of NGS technologies Sanger sequencing required high DNA concentrations, ranging from 10µg to 50µg (Polz and Cavanaugh, 1998). Early NGS technologies were gravitating towards the use of micrograms of DNA, whilst nowadays smaller concentrations are required ranging in nanograms (Hoeijmakers *et al.*, 2011; Marine *et al.*, 2011). The utilisation of smaller amounts of DNA will allow the removal of the DNA concentration steps, reducing both costs and potential sequencing bias.

Table 1: Comparison of main NGS platforms used for marine studies: 454, Illumina and SOLiD (modified from (Glenn, 2011; Liu *et al.*, 2011; Scholz *et al.*, 2012)). Important characteristics for each platform are shown: type of sequencing, difference in amplification protocol, read length and cost per run. These characteristics have to be considered depending on sampling and data structure.

Platform	**454-Roche**	**Illumina-Solexa**	**SOLiD-ABI**
Sequencing method	Synthesis (pyrosequencing)	Synthesis	Ligation
Amplification method	Emulsion PCR	Bridge PCR	Emulsion PCR
Read Length	400- 500 bp (soon 800 bp)	≥100 bp on each end of templates	75 bp
Cost per run	Smaller numbers of middle to extended reads at relatively high cost per Mb of sequence	Larger numbers of short to middle length reads at lower cost per Mb	
Pros	Long reads are more suitable for initial genome and transcriptome characterisation. Improved mapping in repetitive regions.	Lower costs and increased number of reads associated with short read length. Leads in number and % of error-free reads	Two-base encoding, which provides inherent error correction.
Cons	High reagent cost, high error rate in homopolymer repeats	Low multiplexing capability of samples	Not suitable for metagenomic Long run time.
Error type	Indel (insertion or deletion)	Substitution	Indel
Errors rate	For all the platforms errors increase near the end of maximum read length.		

Two different approaches are used in NGS based studies, either an amplicon-based or shotgun sequencing-based approach (Mineta and Gojobori, 2016) (Figure 1.5). Finally, metagenome shotgun sequencing refers to sequence data sampled from the environment, with the term metagenome used for the first time in 1998 (Handelsman *et al.*, 1998).

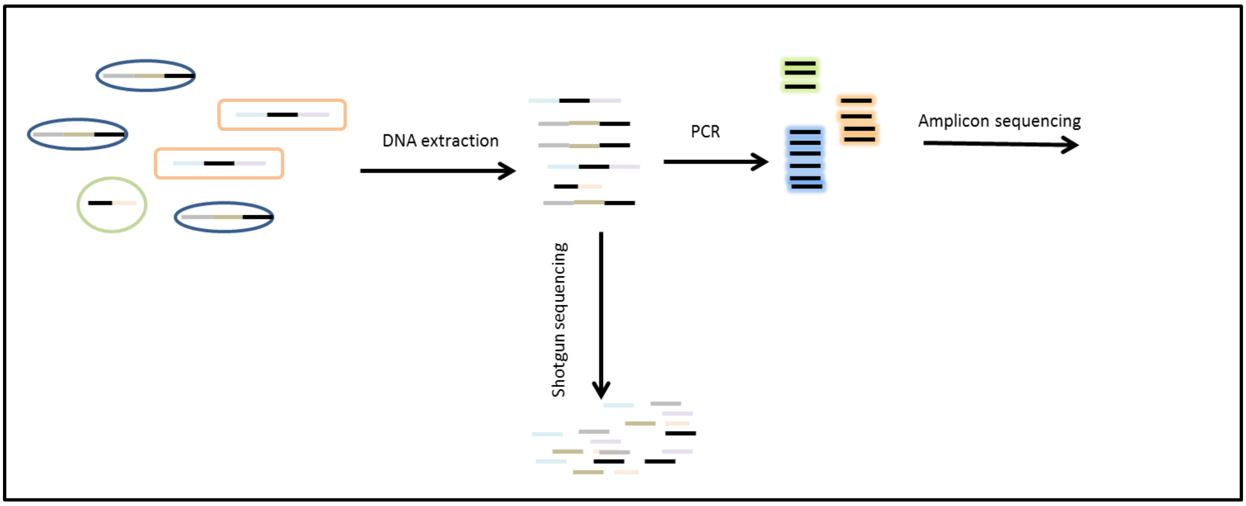


Figure1.5: Schematics of Amplicon sequencing versus shotgun sequencing. Black lines represent conserved sequences such as 16S and 18S that can be utilised for taxonomic identification.

Due to the presence of conserved genes in both prokaryotes and eukaryotes, these organisms can be studied utilising an amplicon-based approach (Figure 1.5). In this method Polymerase Chain Reaction (PCR) is employed to amplify the 16S rRNA gene for prokaryotes (Woese and Fox, 1977; Pace, 1997) and 18S rRNA gene for eukaryotes (Stoeck *et al.*, 2010) which represent the most common molecular markers used for the respective groups. Differences between these conserved regions are then utilised to distinguish between different groups. This method is linked to a barcoding approach (Valentini *et al.*, 2009) based on the small subunit of the rRNA gene similarities in which microbial species correspond to "Operational taxonomic units" or OTUs (Olsen *et al.*, 1986). Difficulties with this method still exist including the need for universal primers, which are still lacking especially for microbial eukaryotes (Stoeck *et al.*, 2010).

The second, shotgun sequencing-based, approach has proven fundamental to the study of microbes and specifically of viruses due to the lack of conserved genes within this group that are convenient for PCR amplification. A large number of short sequences are produced through this method but, unlike the amplicon-based approach, they derive from different regions of the genome (Mineta and Gojobori, 2016) (Figure 1.5). These fragments can be bioinformatically assembled and reference databases can be utilised to look for homologous regions (Kunin *et al.*, 2008; Thomas *et al.*, 2012).

1.9 Thesis outline

Microorganisms are known to form complex ecological interactions and not to survive as isolated cells (Faust and Raes, 2012); these interactions shape key ecological and biogeochemical processes. Thanks to the advancement in NGS technology we are now able to study these organisms, which are invisible at the naked eye and are difficult to study using purely culture based methods (Hugenholtz, 2002). Despite global efforts to study the microbiome, the majority of studies don't address these communities as a whole (Zinger *et al.*, 2011). Throughout this study an alternative and innovative approach is proposed to study microbial diversity in all its complexity, allowing the detection of the most abundant phylotypes. This method can be easily implemented in time series monitoring of the marine environment, opening the door to a more integrated approach of oceanographic sampling, thereby allowing for better parameterisation of global biological models. The inability to characterise microbial assembalges through visual identification has created a drawback in marine monitoring (Goodwin et al 2017). The techniques and methodologies utilised throughout this study will show the possibility of a cost efficient approach that can be used to exploit ecosystem integrated monitoring. The identification of marine microbes through genetic characterisation using smaller volumes of water will hopefully allow the use of microbial data to assess properly marine ecologica status with proper integrated monitoring.

In the first results chapter, questions of experimental design for 16S rRNA gene NGS projects will be considered together with their implications for downstream analyses. To reach a representative variety of environments nine sampling stations, representing both coastal and open ocean environments in northern and southern hemisphere latitudes, will be analysed. In this chapter the use of three replicates obtained through PCR amplification of the prokaryotes V4 region of the 16S rRNA gene will be exploited to address the characterisation of the most dominant phylotypes in environmental samples. The replication approach together with the removal of singletons (OTUs presents with a single sequence) will add robustness to the analysis. Throughout this first results chapter it will be confirmed the robustness of the replication strategy by using rarefaction analyses in combination with subsampling at varying sequence depths.

The second results chapter addresses the analysis of the oceanic “microbiome” and it’s characterisation as a multi-phylotype community of microbes, which in the aquatic environments range from the numerically dominant viruses to the ecologically important and diverse climate-regulating phylotypes of unicellular phytoplankton. The recent advances in NGS are starting to reveal the diversity and biological complexity of marine microbes. Here results derived from sampling at one station are used to develop a bioinformatics pipeline and test different thresholds to remove sequencing bias; furthermore, the hypothesis that a small volume of water can be utilised to evaluate the most abundant fraction of the microbial community will be tested.

In the last results chapter, the hypothesis that “everything is everywhere, but the environment selects” and the subsequent conclusion of the absence of marine barriers is addressed. Amplicon sequencing was utilised to characterise the host fraction (prokaryotes - V4 region of the 16S rRNA gene and eukaryotes – V9 region of the 18S rRNA gene), whilst metagenome shotgun sequencing was used to analyse the viral fraction. The samples, collected from six stations situated in the south-east Indian Ocean, south-west Indian Ocean and Southern Ocean, will be used to characterise the most abundant phylotypes above and below the Antarctic Polar Front (APF). These sampling choices allowed the testing of the hypothesis that ocean fronts can act as an open ocean barrier for the microbial community.

The overarching objective of this project aims to bring a new insight on the study of marine microbes, providing a new monitoring tool to keep track of changes in microbial communities due to natural occurring events as well as human induced phenomena. Microbial communities, from virus to protists, are described from six stations from the Southern Indian Ocean and Southern Ocean. Furthermore this study brings a new understanding on the role of “invisible” marine barriers, providing a step towards the understanding of the role of microbes in the oceans.

Chapter 2: PCR amplification replicates and singleton removal in tag amplicon NGS projects: a method for the removal of erroneous diversity

2.1 Introduction

Sampling of microbial populations across the globe has become a widespread activity, and projects such as the Earth Microbiome Project (Gilbert *et al.*, 2011), the International Census of Marine Microbes (Amaral-Zettler *et al.*, 2010), the *Tara* Expeditions (Bork *et al.*, 2015) and the Micro B3 led Ocean Sampling Day events (Kopf *et al.*, 2015) provide protocols to sample and compare microbial community diversity via next generation amplicon sequencing. The first studies using this technology were based on the 454 pyrosequencing technology, but more recently Illumina amplicon HiSeq and MiSeq platforms have become popular, yielding increased output (albeit shorter reads) for reduced cost (Caporaso *et al.*, 2011). A recent evaluation (Caporaso *et al.*, 2012a) showed that both Illumina platforms are effective for capturing and exploring microbial populations and nowadays these techniques are widely used to explore microbial diversity in both marine and terrestrial environments (Caporaso *et al.*, 2011, 2012a, Gilbert *et al.*, 2012, 2014).

The experimental design of a next generation sequencing (NGS) study targeting microbial diversity is crucial for determining the level of diversity potentially captured and characterised, as well as contributing to the confidence with which findings can be reported. Specifically, the amount of water filtered to extract DNA, the sequencing technology and sequence depth (i.e. how many times a specific nucleotide is present, on average, in the raw data, Sims *et al.*, 2014) all significantly influence the results (Zinger *et al.*, 2012; Ghiglione *et al.*, 2005). In order to capture marine microbial diversity, several nested issues of scale need to be addressed. First, given a specific locality, how much water needs to be filtered in order for the extracted DNA to be representative of the microbial diversity in that locality at a given time point? Second, how much of the extracted DNA needs to be sequenced to capture the microbial diversity in the present sample? And third, in the case of amplicon sequencing, how many sequence reads are required in order to adequately capture the microbial diversity in the sample? Hereinafter these nested issues of scale will be referred as: SC1, SC2 and SC3 for the first, second and third, respectively.

The answers to these questions depend on the 'species' abundance distribution(s) across marine waters, i.e. how many species, or taxa, in a community are present across the range of abundances. Typically there are many taxa present at low abundance (i.e. "rare") with few taxa present in higher proportions ("dominant"); this phenomenon can be captured by a range of different species abundance distributions (Gilbert *et al.*, 2012; Huber *et al.*, 2007). Given such theoretical constraints, all diversity need not be captured in order to estimate the degree of diversity (Curtis *et al.*, 2002), although these extrapolation methods cannot evaluate fully the communities.

As sequencing read depth increases, the number of DNA strands sequenced increases, yet so does the number of errors introduced by sequencing (Nakamura *et al.*, 2011). Furthermore during the sequencing process saturation will be reached, meaning that every strand of DNA in the subsample has been captured. This saturation can be assessed via rarefaction analysis. When saturation is reached, the rate of increase of new sequences observed as more sequences are generated begins to plateau. Failure to reach saturation may arise either if the read depth is not sufficient to cover the range of DNA sequenced (SC3), or if the sequenced DNA is not fully representative of the diversity of the sample's DNA content (SC2) or of the region from which the sample was taken (SC1).

Following a similar logic, the effects of adding PCR replicates on maximising the level of saturation in the number of taxa or operational taxonomic units (OTUs based on 97%

sequence identity) or phylotypes (i.e. taxonomic assignments) observed can also be considered. If a single PCR replicate is enough to adequately characterise the sample, then very few new OTUs would be expected to be found when additional PCRs are performed. Nevertheless, a recent eukaryote study showed that PCR replication can significantly increase the number of OTUs detected (Schmidt *et al.*, 2013). The presence of new OTUs due to replications may reflect real variation of the community sampled, as well as indicating that the DNA aliquot used in a single subsample PCR is insufficient to describe the diversity of the extracted DNA pool. Furthermore, the use of replication helps to identify errors associated with PCR amplification or sequencing. While distinguishing errors from real variation is a difficult process, there is more confidence in the OTUs identified if these are present in more than one replicate PCR.

In this study, the first scaling issue of water quantity (SC1) will not be addressed, because it has been addressed in previous studies (Ghiglione *et al.*, 2005; Dorigo *et al.*, 2006). Therefore the ultimate bacterial community diversity is not likely to be comprehensive. Nonetheless, it should give an indication of what is dominant in the water column in each relative volume of water at that point in time. Here the focus will be based on addressing scaling issues SC2 and SC3 using a triplicate independent PCR design (i.e. for each biological sample the DNA was extracted and subsequently three independent PCRs amplifications were performed) and high depth Illumina single end reads. Single reads provide similar estimates of biodiversity as paired end reads (Caporaso *et al.*, 2011). The sampling of six different environments, from costal to open ocean, will add robustness to the study. Through comparing PCR replicates it is possible to identify overall differences (e.g. if one of the PCRs is significantly different to others with respect to the number of common/unique OTUs) and also get a sense of which level of diversity can be captured with confidence. Specifically, it is possible to report which OTUs are likely to be observed across all PCRs and in what abundance. In addition, a sample that was exposed to a treatment regime to simulate future high CO_2 scenario was embedded in this dataset. Given that this manipulation was not replicated in both space and time, it was not possible to draw any meaningful conclusions on the significance of the changes observed, however it will nonetheless indicate whether sample preparation, manipulation or perturbation has an effect on downstream analysis and thus diversity predictions. Finally, I will look at possible variation of taxonomic annotations to ensure absence of variation due to sequencing depth through rarefaction analysis.

2.2 Materials and methods

2.2.1 Sample Preparation

A total of ten samples were analysed in this study. Seven samples (stations S1-S6, Figure 2.1a, Supplementary Table 1) were collected during the second cruise (RR1202; Feb-Mar 2012) of the project "The Great Southern Coccolithophore Belt" on board of the research vessel (R/V) Roger Revelle (Scripps Institution of Oceanography). A further oceanic sample (station S9) was collected during the second cruise for the UK Ocean Acidification research program (http://www.surfaceoa.org.uk) on board of the RSS James Clark Ross (JCR271; June-July 2012) as part of the project on Arctic Ocean Acidification. For all oceanic samples, one litre of water was gathered from conductivity-temperature-depth (CTD) rosette sampler at the chlorophyll maximum, and an aliquot of 250ml of seawater was filtered through a 0.45µm polycarbonate filter. The filter was used for the DNA extraction on-board the R/V Roger Revelle and RRS James Clark Ross using the Qiagen DNeasy Blood and Tissue protocol (QIAGEN, Valencia, CA, USA). In addition, two coastal samples (stations S7 & S8, Figure 2.1 and Suppl. Table 1) were collected by gathering surface seawater off a small boat with an acid washed bucket and passing 200ml of 200µm pre-filtered water through 0.45µm polycarbonate filters. Filters were preserved in molecular grade ethanol, stored at 4°C and DNA was extracted in the lab using Qiagen DNeasy Blood and Tissue kit protocol (QIAGEN, Valencia, CA, USA).

To one of the samples (station S1b) a future 2100 climatic scenario was simulated by bubbling CO_2 through a stone, using calibrated gases in order to reach a final CO_2 level of 770ppmv in a temperature controlled incubator at 8°C on deck of the R/V Roger Revelle. The sample was exposed to this condition for 96 hours before the DNA was extracted following the Qiagen DNeasy Blood and Tissue kit protocol (QIAGEN, Valencia, CA, USA).

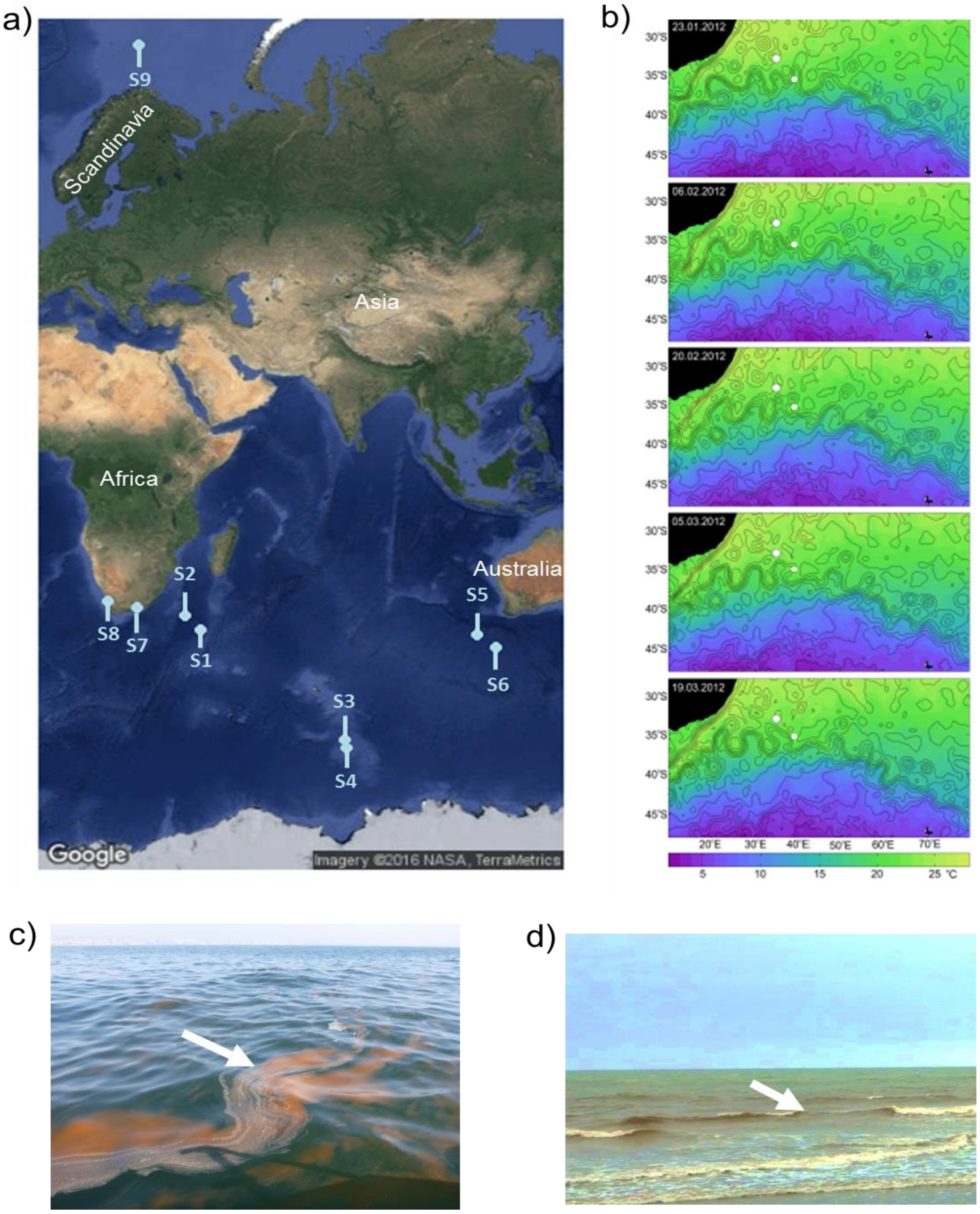


Figure 2.1: Description of the sampling stations. A) Map depicting sample stations (S1 to S9) locations. B) Sea surface temperature overlaid with absolute dynamic topography (ADT). Contours that range from -1 m to 1.4 m in 0.2 m intervals in the region of sampling stations S1 and S2 (the white filled in circle) are shown. The Agulhas Return Current (ARC) is visible as the band of tight contours in ADT that meanders along ≈35°. The images show two months prior to the sample collection (end December-February). C) Red tide sampled in Nelson Mandela Bay, South Africa (station S7). D) Red tide sampled off the coast of Elands Bay, South Africa (station S8).

2.2.2 PCR amplification and preparation for Illumina sequencing

Primers 515F/806R (Appendix I) (Caporaso *et al.*, 2012a, 2011) were used to amplify the V4 region of the 16S rRNA. Primers contained an upstream Illumina adaptor sequence, barcode and linker sequence (5'-3') with three reverse primer constructs designed with unique barcodes. PCR was subsequently performed as followed: 1 to 5µl of the environmental DNA (DNA concentration range from 1.47 to 32.51 ng/µl), to 5X Colourless GoTaq Flexi Buffer (Promega), 1.5µl $MgCl_2$ Solution 25mM (Promega), 2.5µl dNTPs (10mM final concentration, Promega), 1µl Evagreen Dye 20X (Biotium), 0.1µl GoTaq DNA Polymerase (5u/µl- Promega) and 12.9µl of sterile water for a final volume of 25µl for each reaction. This was done to determine the mid-exponential threshold of each reaction, which were run on a Corbett Rotor-Gene™ 6000 (QIAGEN, Valencia, CA, USA). The real time PCR proceeded with an initial denaturation at 94 °C for 3 minutes, followed by 40 cycles of a three step PCR: the cycles were 94°C for 45 seconds, 50°C for 60 seconds and 72°C for 90 seconds.

To determine the number of cycles for the optimal PCR protocol, the logarithmic stage of the reaction was identified by adding SYBR Green (Thermo Fisher Scientific Inc) and tracking the PCR reaction. PCRs were then undertaken in triplicate for each primer pair, in the absence of the nucleic acid SYBR Green stain, and the reactions were stopped at mid-logarithmic stage. PCR products were gel verified, excised from the gel and recovered from the agarose using the Zymoclean gel DNA recovery kit (Zymo Research) according to manufacturer's instructions. Purified PCR products were quantified on the Agilent 2100 Bioanalyser (Agilent Technologies) using the Agilent DNA 12000 kit and were sent to the University of Exeter sequencing facility where the triplicates were pooled at equimolar concentrations and run on the HiSeq 2000 Illumina sequencer. The raw sequences are available at the European Nucleotide Archive (ENA) under accession number PRJEB16346.

2.2.3 Bioinformatic workflow

The workflow for the nine samples collected is shown in Figure 2.2. Analyses were performed using the Bio-Linux 8 system at the Marine Biological Association of the UK. The quality of the raw reads, and later of the processed reads, was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The first and final 10 bases of each read were trimmed to remove non-variable nucleotides and nucleotides called with very low quality score. After trimming, the reads were filtered based on quality scores, only retaining those with ≥ 95% of nucleotide positions called with quality score greater than 20. Trimming and filtering was done using the fastx tool kit (http://hannonlab.cshl.edu/fastx toolkit/).

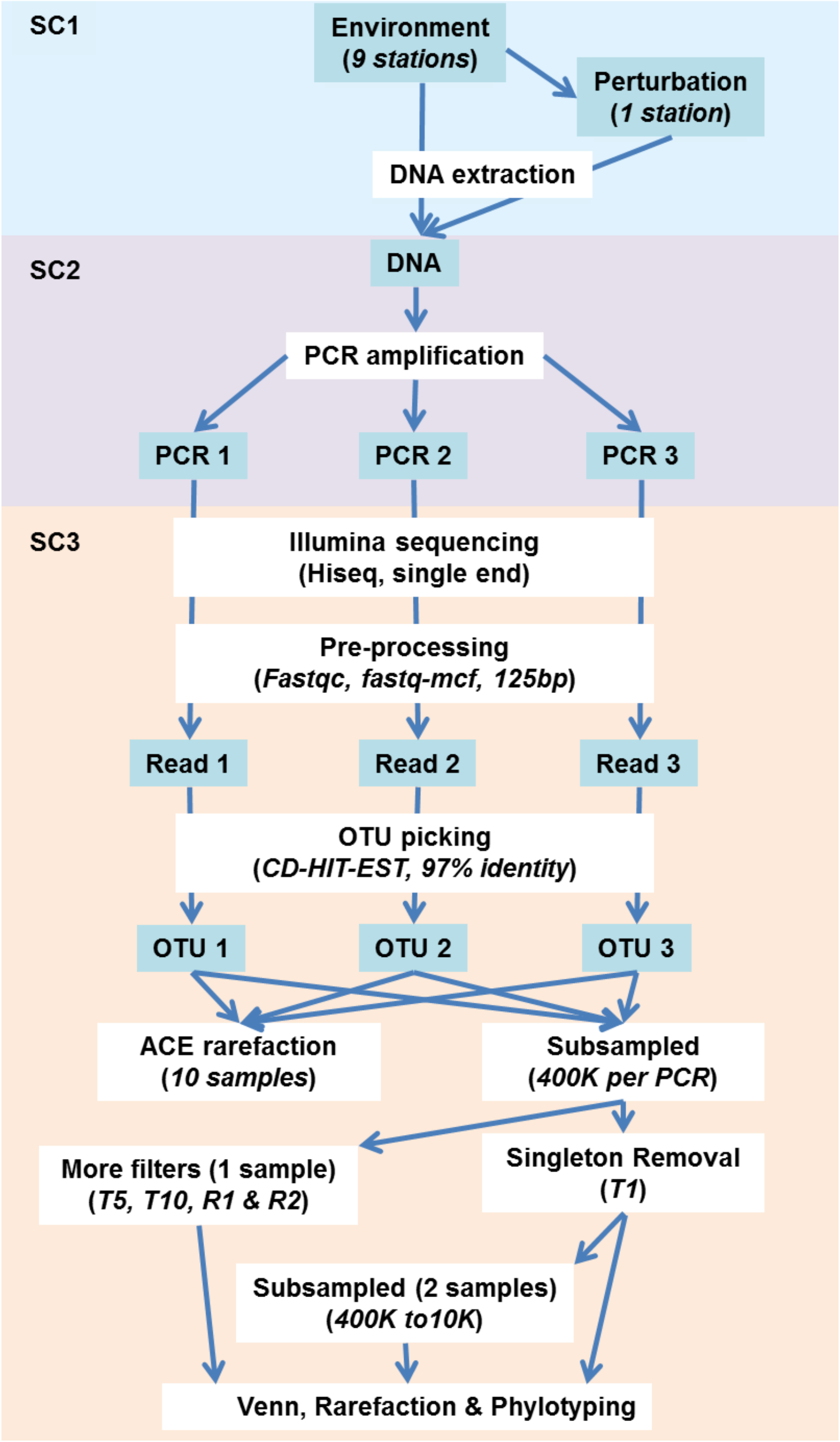

SC1
Environment (*9 stations*)
Perturbation (*1 station*)
DNA extraction
SC2
DNA
PCR amplification
PCR 1
PCR 2
PCR 3
SC3
Illumina sequencing (Hiseq, single end)
Pre-processing (*Fastqc, fastq-mcf, 125bp*)
Read 1
Read 2
Read 3
OTU picking (*CD-HIT-EST, 97% identity*)
OTU 1
OTU 2
OTU 3
ACE rarefaction (*10 samples*)
Subsampled (*400K per PCR*)
More filters (1 sample) (*T5, T10, R1 & R2*)
Singleton Removal (*T1*)
Subsampled (2 samples) (*400K to10K*)
Venn, Rarefaction & Phylotyping

Figure 2.2: Overview of the study workflow. T1: singleton removal; T5: filter removing OTUs observed with a total abundance <5; T10: filter removing OTUs observed with a total abundance <10. R1: filter retaining OTUs observed in at least two independent PCRs; R2: filter retaining OTUs observed in all three independent PCRs.

2.2.4 Defining, subsampling and filtering OTUs

The OTUs were defined using the CD-HIT-EST open OTU picking method in Qiime (Li and Godzik, 2006). This method is based on a similarity threshold rather than a reference database by grouping sequences into clusters so that sequences assigned to each cluster present 97% sequence similarity. The most abundant sequence in each OTU was selected as the representative sequence for that OTU. Both these steps were performed using Qiime[1.8] (Caporaso *et al.*, 2010) using the commands pick_otus.py and pick_rep_ set.py respectively. In order to make comparisons across all the PCRs, the reads from each sample (and replicate) were subsampled down to the lowest read count observed (1.2 million reads). This step was replicated 100 times and the average read count was utilised in the OTU table. Average read counts below one were set to zero and referred to as subsampled OTU (T0p).

OTU filtering was performed on the subsampled OTUs defined by 97% identity, i.e. prior to taxonomic assignment, allowing the direct comparison of the three PCR replicates present for each sample. A first filter to exploit the triplicate design consisted of the removal of singletons, meaning that the OTUs observed in only one of the three replicates with only one read were removed and the code T1 was assigned to this filter. Additionally two further filters were considered to exploit the triplicate design: replicate filter one (R1), which retains OTUs observed in at least two of the three independent PCRs and replicate filter two (R2), which retains OTUs observed in all three independent PCRs. These three filters were compared with the following two filters which, contrarily from the first three, act on the total data set, i.e. totalling the reads across all the PCRs: filter T5 that removes OTUs observed with total abundance less than five reads and filter T10, which removes OTUs observed with total abundance less than ten reads.

2.2.5 Taxonomy assignment

Taxonomy assignments were made on the representative Operational Taxonomic Unit (OTU) sequences using Basic Local Alignment Search Tool (BLAST, Qiime implementation) when ≥ 90% of the target sequence matches the database (SILVA release 119 (Pruesse *et al.*, 2007)) with a BLAST e-value of < 10^{-5}. The community composition, as characterised by the SILVA taxonomy assignments, was visualised based on the relative abundances of the main taxa using the R package ggplot2_2.1.0.

2.2.6 Rarefaction analyses

Rarefaction analyses on sequence depth were performed using Qiime by random subsampling OTU tables 100 times at each sequence depth. The averages of the 100 subsamples were then used to plot the rarefaction curves. Alpha diversity was defined by using the Qiime script alpha_diversity.py for both observed species and ACE indexes using the total OTU count. In addition, rarefaction by PCR was performed to consider the extra diversity captured by performing independent PCRs. All six permutations of the three independent PCRs namely {1,2,3}, {1,3,2}, {2,1,3}, {2,3,1}, {3,2,1}, {3,1,2} were considered and the total number of OTUs observed overall was recorded as each consecutive PCR was added.

2.2.7 Sea Surface Temperature and Mesoscale Circulation

Daily maps of absolute dynamic topography and sea surface temperature were used to examine the mesoscale circulation of the southern hemisphere oceanic regions in the six months prior to sampling at the station. Images for Figure 2.1.b were selected from the two months period prior to sampling at intervals of two weeks. The absolute dynamic topography fields were calculated by Aviso at 1/4 degree horizontal resolution from all the remotely-sensed altimetry mission data available at a given time referenced to a 20 years interval (Rio *et al.*, 2013). High resolution (1/20 degree) sea surface temperature data was produced from the Operational Sea surface Temperature and Ice Analysis (OSTIA) system using both in situ and satellite data (Donlon *et al.*, 2012).

2.3 Results

The bioinformatic pipeline, for the Illumina HiSeq single-end reads, involved pre-processing of raw reads, OTU picking, taxonomic assignment (phylotyping) and rarefaction analyses (Figure 2.2). The cleaned subsampled reads of the V4 16S rDNA gene region from all nine stations were used for the analyses (Table 2.1, Figure 2.1 & Supplementary Table 1). In addition, a second sample (S1b) was also taken at station S1; this sample was incubated for four days under a future high pCO_2 scenario condition. This sample was subjected to the same bioinformatics analysis as the other nine samples.

The nine sample stations represent both open ocean (S1-S6 & S9) and coastal (S7 & S8) environments, where the open ocean samples were collected at various depths (5 to 60 m) at the deep chlorophyll maximum layer. Measurements of absolute dynamic topography and sea surface temperature showed that the Agulhas Return or Antarctic Circumpolar Currents did not directly influence stations S1 and S2 during the time of sampling (Figure 2.1b). Station S9 is the most northern station (~400km miles north of Scandinavian Peninsula), while stations S3 and S4 are located in the Southern Ocean (~1000km north of Antarctica). The two coastal samples, stations S7 and S8, were collected at a time when a red tide algal bloom events occurred off the east and west coast of South Africa, respectively (Figure 2.1c & 2.1d, Supplementary Table 1).

Table 2.1: Number of raw and cleaned reads within replicate PCRs

Sample	Replicate	Raw reads	Pre-processing and QC	Reads of equal length (125 bp)	Final fraction Raw count (%)
S1a	1	1,331,542	773,343	741,033	55.65
	2	1,695,911	1,161,634	1,117,576	65.90
	3	1,626,930	863,867	841,639	51.73
S1b	1	3,255,784	2,886,069	2,374,540	72.93
	2	1,945,967	1,716,341	1,393,673	71.62
	3	2,892,668	2,592,385	2,176,762	75.25
S2	1	983,760	443,622	437,790	44.50
	2	1,458,024	627,698	619,255	42.47
	3	1,550,314	646,303	637,701	41.13
S3	1	1,491,664	622,030	609,658	40.87
	2	1,409,872	795,524	781,413	55.42
	3	1,754,942	878,836	864,910	49.28
S4	1	974,224	438,389	434,686	44.62
	2	1,609,312	721,401	714,793	44.42
	3	1,468,624	795,217	788,622	53.70
S5	1	1,497,998	805,139	785,754	52.45
	2	838,777	725,672	706,520	84.23
	3	1,253,530	725,301	708,433	56.52
S6	1	1,477,596	664,590	659,890	44.66
	2	1,695,898	761,187	755,509	44.55
	3	771,891	696,673	691,459	89.58
S7	1	1,399,938	657,317	636,419	45.46
	2	1,333,355	727,012	705,914	52.94
	3	1,044,699	496,969	484,443	46.37
S8	1	1,392,915	883,292	878,091	63.04
	2	1,529,536	891,374	879,127	57.48
	3	1,258,768	819,381	814,317	64.69
S9	1	2,192,158	563,355	561,702	25.62
	2	1,184,300	524,804	522,963	44.16
	3	1,238,172	645,438	642,442	51.89

2.3.1 Defining OTUs

The open picking OTU algorithm CD-HIT-EST with 97% sequence identity was chosen to ensure that OTUs are defined independently of a reference sequences and to provide reproducibility. The total number of reads (after pre-processing, quality control, and normalisation to equal length) ranged between 434,686 and 2.37 million (Table 2.1). For purposes of comparison, the samples were normalised by randomly subsampling 1.2 million reads per site (400,000 per replicate), which clustered into a range spanning from 10,486 to 22,452 OTUs (Table 2.2). The OTUs in the high pCO_2 treated sample, S1b, showed a reduction in overall biodiversity compared to the untreated control, S1a (18,215 to 12,428 OTUs: down by 32%). A total of 5,767 OTUs which differed from S1a were recorded, whilst almost 30% of the sequences were maintained in the high pCO_2 treated sample (Figure 2.3).

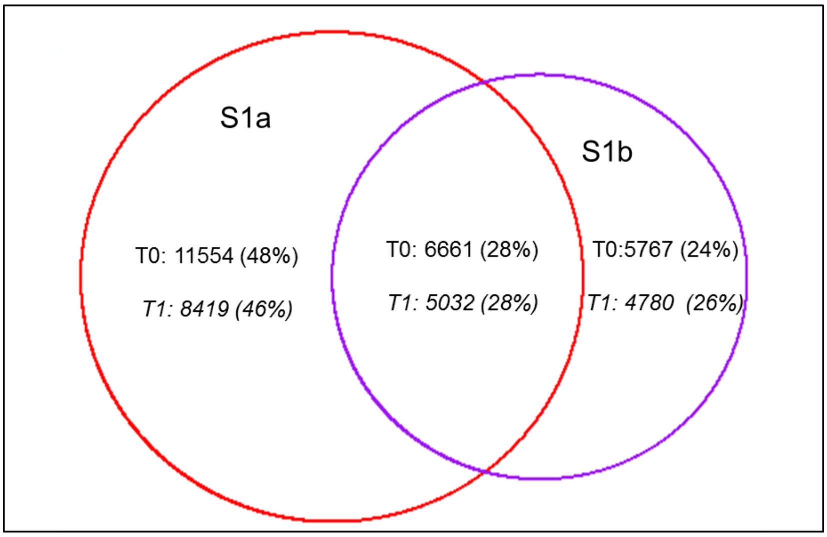


Figure 2.3: Venn diagram comparing OTUs in S1a versus S1b when no filter is applied (T0p) and after use of filter T1; percentages values in brackets.

Table 2.2: The number of OTUs and reads before and after filtering. T0p was defined at 97% sequence identity when the final reads from each of the independent PCRs is randomly subsampled (400,000 reads), pooled (1.2 million reads) and clustered; the PCR label was retained to allow for comparison and subsampling of this OTU table within PCR in subsequent analyses. T1 refers to the removal and clustering of singletons. Rep1, Rep2 and Rep3 refers to singletons found in only one, any two or all three replicates, respectively.

Samples	OTUs (T0p)	OTUs (T1)	% OTUs removed (T1)	Reads (T1)	% Reads retained (T1)	Reads removed (T1)	Reads Rep1 [%]	Reads Rep2 [%]	Reads Rep3 [%]
S1a	18,215	13,451	26	1,193,997	99.553	5,361	4,203 [78]	1,050 [20]	108 [2]
S1b	12,428	9,812	21	1,196,163	99.730	3,234	2,079 [64]	912 [28]	243 [8]
S2	18,203	13,172	28	1,193,795	99.549	5,414	4,648 [86]	766 [14]	0 [0]
S3	15,877	11,093	30	1,193,675	99.564	5,222	4,387 [84]	712 [14]	123 [2]
S4	15,603	11,909	24	1,196,945	99.673	3,929	3,459 [88]	470 [12]	0 [0]
S5	22,452	15,543	31	1,191,624	99.341	7,907	5,998 [76]	1,648 [21]	261 [3]
S6	13,675	9,006	34	1,194,028	99.548	5,426	3,994 [74]	1,186 [22]	246 [4]
S7	10,486	8,124	23	1,197,318	99.787	2,555	2,169 [85]	386 [15]	0 [0]
S8	15,808	12,177	23	1,195,330	99.680	3,834	3,434 [90]	382 [10]	18 [0]
S9	14,466	11,404	21	1,195,744	99.726	3,283	2,851 [87]	402 [12]	30 [1]

2.3.2 Comparison of numbers of OTUs observed in independent subsampled PCRs

Only between 15 and 35% of OTUs were observed across all three independent PCRs for the nine sampling stations (Figure 2.4a); however, the reads in these shared OTUs accounted for between 84 and 98% of the total 1.2 million subsampled reads per station (Table 2.2 & Figure 2.4b). Only the manipulated sample S1b showed higher OTU counts across all three PCRs (54%) but with a similar read dominance of 99% (Figure 2.4a). Notably both OTU richness and read counts across all three PCRs increased when compared to sample S1a (Figure 2.4a & 2.4b). A closer examination of the read counts for the OTUs present in any one of the three independent PCRs, showed that more than 90% of the OTUs had a read count below five, and 95% had a read count smaller than eight (Figure 2.5). Similarly, the read counts for OTUs present in only one of the three PCRs accounted for less than 8% of the total reads despite making up between 45% and 65% of the OTUs (Table 2.2). Removal of the singletons (T1 filtering) in either one, two or across all three replicates resulted in the loss of between 2,555 and 7,907 reads from the subsampled 1.2 million reads (Table 2.2); this led to the reduction in the overall number of OTUs from a minimum of 21% in station S1b to a maximum of 34% in station S6 (Table 2.2). Furthermore, the majority of the singletons, between 64% and 89.5%, were observed in only one of the three PCR replicates. The application of this filter reduces the OTUs count from a rage between 10,486 and 22,452 to a range between 8,124 and 15,543 OTUs (Table 2.2).

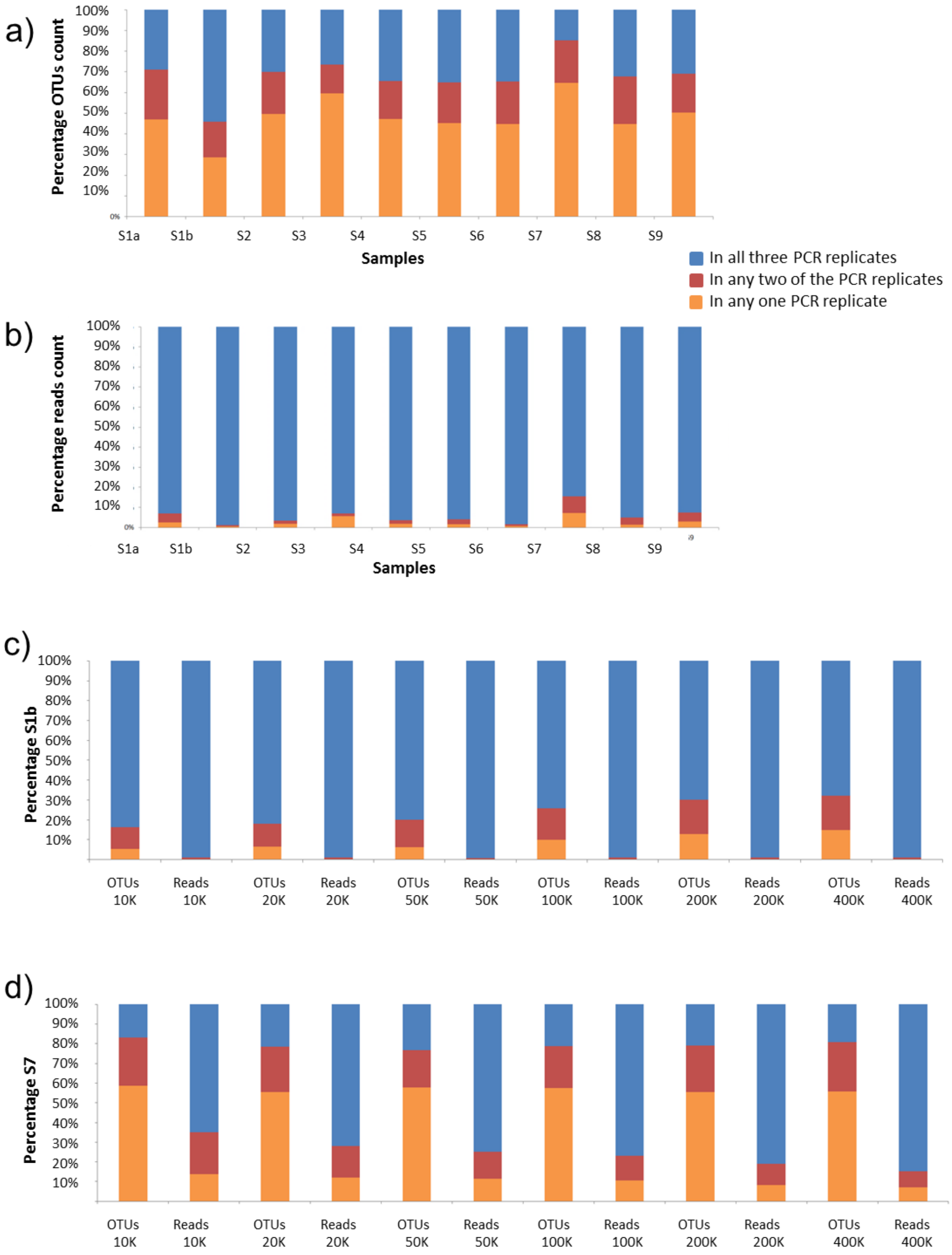


Figure 2.4: Reads and OTUs distribution across replicate PCRs. a) Percentage of OTUs present in one, two or three replicates for each sample; b) percentage of reads present in one, two or three replicates for each sample; c) read and OTUs at different sequence depth after T1 (removal of singletons) for sample S1b; d) Read and OTUs at different sequence depth after T1 (removal of singletons) for sample S7

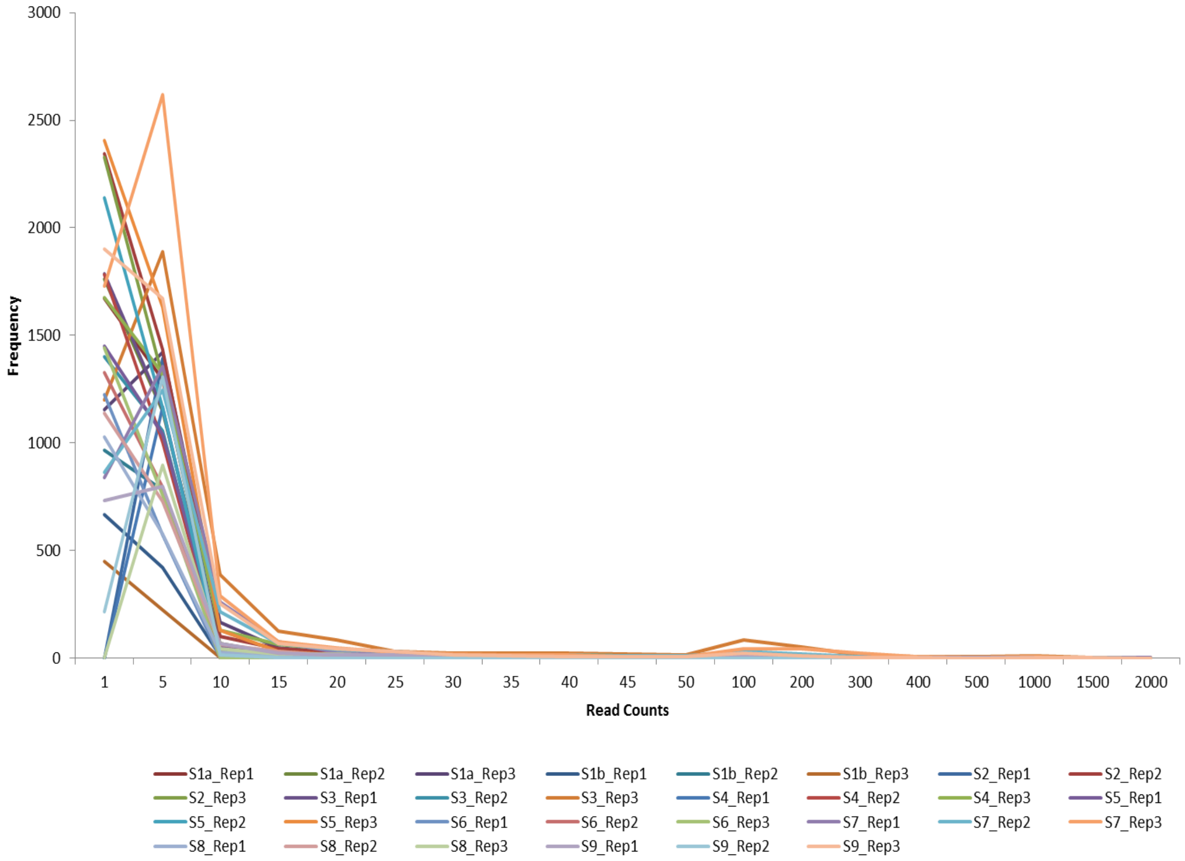


Figure 2.5: Frequency of read counts per replicate PCR when OTUs are present in a single PCR.

2.3.3 Impact of sequencing depth on OTU distribution

Further subsampling of sample S1b, in which the highest OTU count across all replicate PCRs could be observed (Figure 2.4a), even when singletons were removed, revealed a relative decrease in OTUs across all three replicates from 84% to 68% with increasing in sequencing depth (blue bar Figure 2.4c). This was, however, not observed for S7 where the OTUs common across all replicate PCRs increased from 17% to 19% with increasing sequence depth (Figure 2.4d).

Rarefaction analyses performed using an abundance-based coverage estimator (ACE), to further assess sequencing depth by subsampling each independent PCR, showed that the variation between PCRs was greater in some samples (Figure 2.6). For example, station S7 showed the greatest variation across replicate PCRs whilst station S1b the lowest. Lower variation within S1b was observed compared to sample S1a. In addition ACE predicts S5 to have 50,000 OTUs, while S8 and S1b to asymptote at around 25,000 OTUs (Figure 2.6).

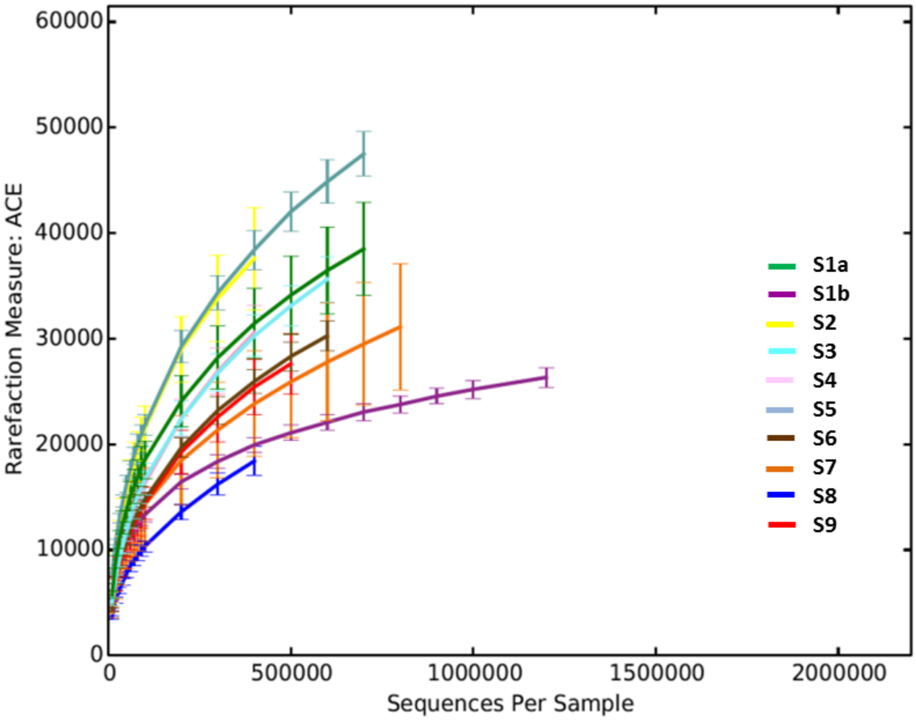


Figure 2.6: Rarefaction on the subsampled OTU table. Aggregating reads across independent PCRs for all samples.

To further address SC2 (i.e. the sequence depth for at least one sample required to obtain saturation of sequences present in the sample prepared for sequencing), the sample S1b containing the greatest read depth (a total of 5.9 million reads from the independent PCRs, Table 2.1) was further analysed. Standard rarefaction analyses were performed using the OTU tables filtered using all five filtering regimes (Figure 2.7). Results showed that the more stringent the OTU table filter (such as R2 and T10), the lower the sequencing depth requirement, as these OTU tables contain fewer OTUs comprising a small number of reads (Figure 2.7a). To reach saturation for OTUs observed across all independent PCRs, a total sequencing depth of three million reads is required for this sample. Furthermore, saturation began to occur at five million reads with the least stringent filter T1, which removes only singletons. A similar saturation effect should occur in all the other samples; in fact the application of the filters to the remaining nine samples showed a similar OTU collapsing effect occurred across all the samples (Supplementary Table 2).

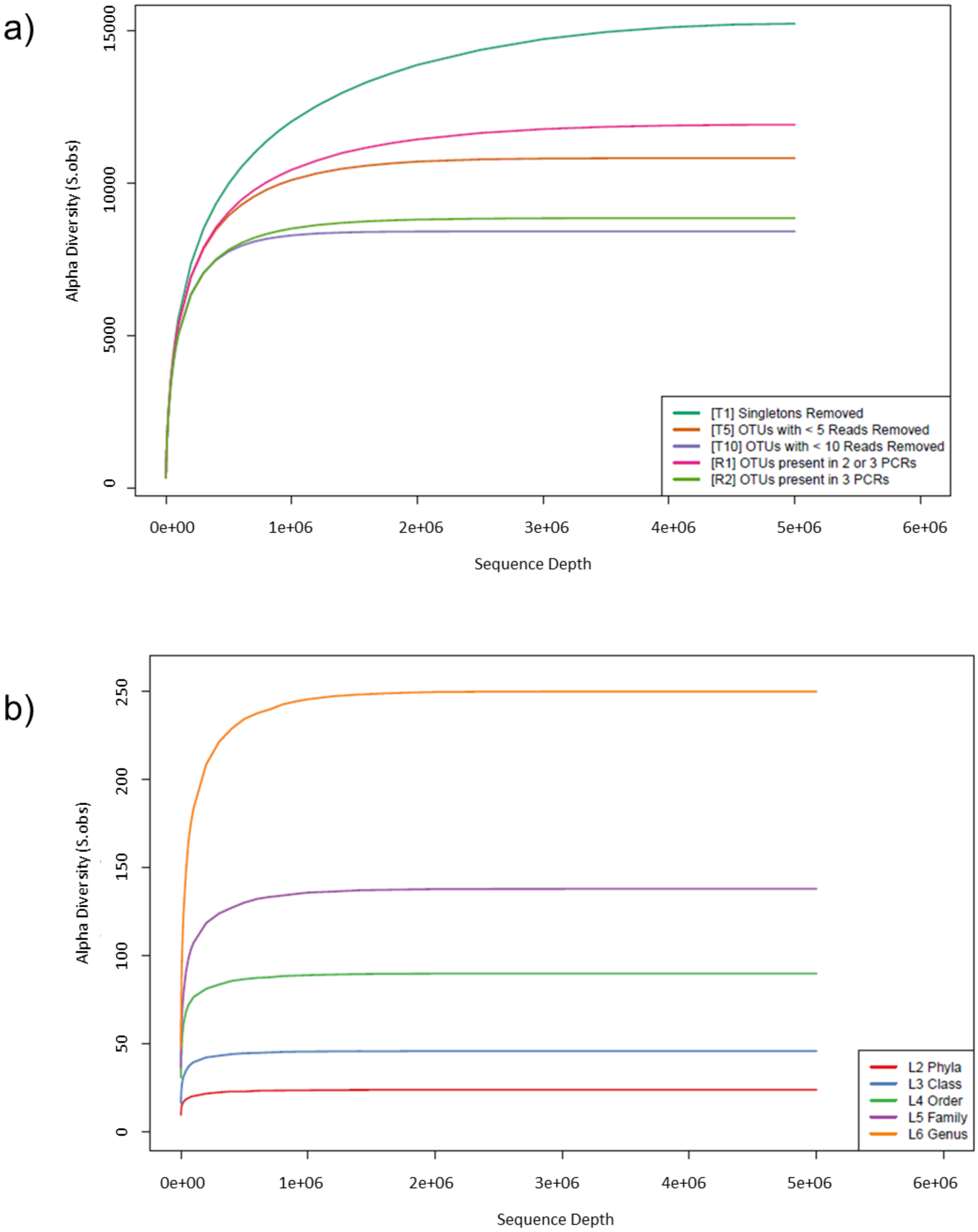


Figure 2.7: Rarefaction for sample S1b on the subsampled OTU table. a) Comparing the range of OTU table filters T1, T5, T10, R1, R2 prior to taxonomic annotation for S1b; b) comparing the impact of different taxonomic levels (assigned using the SILVA 119 annotation) on the OTU table filtered with R2 (removing OTUs not observed across all three PCRs)**.**

2.3.4 Taxonomy Assignment

Assigning taxonomy to the OTUs, and collapsing the dataset in order to combine OTUs assigned to the same taxon reduced the total number of OTUs (Supplementary Tables 2 & 3). For sample S1b, the rarefaction performed on these collapsed OTU tables indicated saturation occurred at fewer reads (Figure 2.7b). The number of reads required for saturation increases with the level of taxonomic assignment from 0.5 million at level 2 of SILVA taxonomy (mainly represented by Phyla) through to two million at level 6 (including predominantly Genera). This analysis suggests that sampling requirements depend on exactly how OTUs are defined. If OTUs are defined without being assigned to phylotypes, then sampling requirements are higher. This was consistent across all samples (Supplementary Tables 2 & 3).

Irrespective of the database selected (data not shown) for taxonomic annotation, the degree of collapse from the OTU table defined by 97% sequence identity is three orders of magnitude down to phyla level (level 2), and two orders of magnitude down to family level (level 5).

Effects of sequencing depth on read and OTU numbers, for samples S1b and S7 (Figure 2.4c & 2.4d), were further analysed to test the effects on the taxonomy annotation. Effects of subsampling (10K-400K) showed for station S1b a decrease in relative abundance of certain taxa such as the cyanobacteria together with the increase of other taxa such as Alphaproteobacteria with the increase of the sequence depth (Figure 2.8a). Whilst for station S7 little variation in phylotypes was observed with increasing sequencing depth across all three independent PCRs (Figure 2.8b). The bacterial phylotyping for all samples obtained independently of replicate PCRs (Figure 2.8c) shows clearly the dominance of cyanobacteria in these south-west Indian Ocean samples (stations S1 & S2), while a variety of proteobacteria lineages dominated in the Southern Ocean (stations S3 & S4) and south-east Indian Ocean samples (S5 & S6). Gammaproteobacteria and Bacteriodetes dominated the two coastal red algal bloom events on the east (station S7) and west coast (station S8) of South Africa, respectively; finally, the most northerly station (S9) has populations similar in composition to the two most southerly stations (Figure 2.8c).

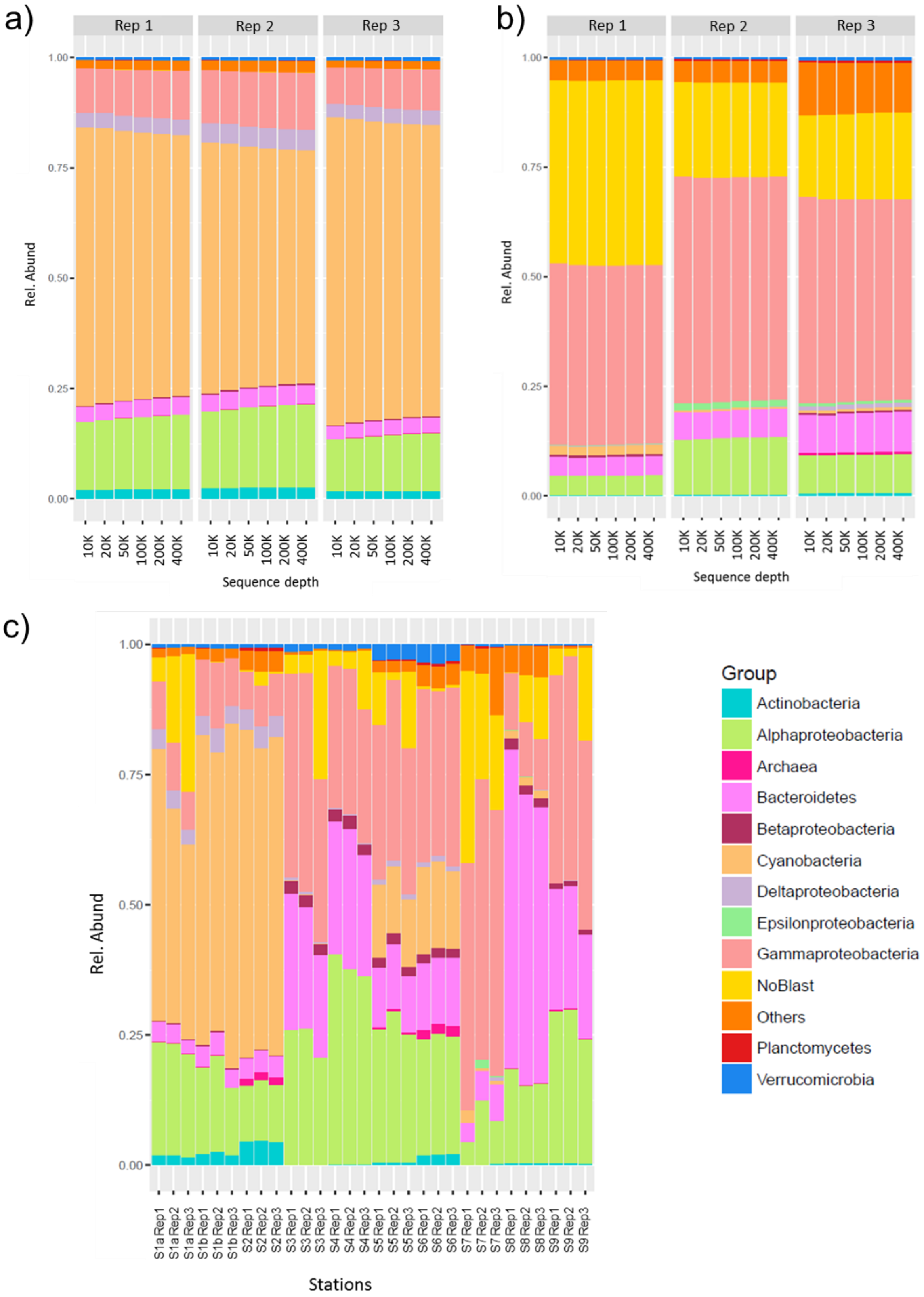


Figure 2.8: Phylotypes composition for all prokaryotic samples: a) sample S1b including the three replicates with different sampling depths; b) sample S7 including the three replicates with different sampling depths and c) all samples

2.4 Discussion

In the introduction, three nested issues of scale were raised; within this study, two scaling issues, SC2 (replicate independent PCRs) and SC3 (the sequence depth for at least one sample required to obtain saturation of sequences present in the sample prepared for sequencing), were addressed. The first scaling issue was addressed in previous studies such as Ghiglione *et al.* (2005) and Dorigo *et al.* (2006) and therefore bypassed in this study. Furthermore, despite the fact that the triplicate PCR design introduces additional stages to the analysis, it allowed to explore robustness using comparative analyses.Due to the smaller sample volume of water collected, levels of saturation were expected to vary from other studies depending on the total amount of DNA extracted. The choice of different type of sampling stations, from costal to open ocean, allowed the testing of the study design and check that the observed patterns are repeatable across different marine environments.

One of the primary aims of this study was to examine robustness and assess how likely a OTU that is observed in a single sequencing experiment would be reproducible with further independent sequencing efforts on PCR products generated from the same extracted DNA. Filtering the OTU table to contain only OTUs common to all three PCRs provided a provided robustness to the analyses and consequently an increase in confidence that each independent PCR replicate was representative of the sample; furthermore it showed how reproducible the results would be if more than one PCR amplification replicate was performed. In fact, when OTUs were defined without the removal of singletons, a higher number of OTUs representing the rare fraction were present; nevertheless, these OTUs were mainly characterised by presence in a single PCR with a single read. Therefore removal of these sequences was consistent with the removal of sequencing or PCR errors. Analysis suggests that multiple replicate PCRs, singleton removal and minimal sequence depth (which is dependent on the sample), provides a good overall representation of the diversity present in the sample. This conclusion was based on the observation that whilst the high percentages of the OTUs defined by 97% sequence identity overall were not common to all three PCRs, these OTUs were contained only in a small fraction of total number of reads. Therefore, the fact that around a third of the OTUs defined at the 97% sequence identity level were common to all three PCRs suggests caution must be taken as to the way rare OTUs are interpreted, particularly when they form part of a comparative study either through time or space. Whilst the singleton removal in presence of multiple PCRs allows a better understanding of the rare fraction, absence of time series made it difficult to establish which portion of these PCR-specific OTUs constitutes error, either sequencing or PCR amplification, and which truly represents a low abundance OTU. Time series information combined with replicate PCR and sequencing design could be used to address this further.

As the primary motivation for the experimental design of this study was to address scale issues SC2 and SC3, the dataset, at this stage, was not exhaustively compared with previous studies. However, observations on the taxonomic community composition for some of the stations were consistent with comparable datasets (Hunt *et al.*, 2013; Zinger *et al.*, 2011). Specifically the dominance of *Prochlorococcus* in the southern Indian Ocean samples was consistent with a recent global study of *Prochlorococcus* abundance (Flombaum *et al.*, 2013). Furthermore, previous studies on marine habitats that sampled the deep ocean (Huber *et al.*, 2007) and a coastal station off the UK (Gilbert *et al.*, 2012) found similar species richness, 18,537 and 8,794 OTUs, respectively, from a sequence depth of around 700,000 reads. This is despite the use of a different 16S rRNA region (V6) and an older pyrosequencing technology.

Interestingly, the high pCO_2 treated sample, S1b, showed a reduction in overall biodiversity compared to the untreated control, S1a. In addition OTUs different from the untreated sample were observed. These differences cannot be assess fully due to absence of a control sample (i.e. incubated for same time and temperature as the treated one). Further experiments will be necessary to confirm the reasons behind gain and loss of OTUs in the high

pCO_2 scenario. At this stage it is unclear what caused the variation in community and requires further verification that could be further assessed in a control mesocosm experiment.

Nevertheless, the origin of these new sequences could be attributed to the changes in the environment. Climate change has been shown to induce numerous shifts in the distributions and abundances of species (Brun *et al.*, 2016; Barton *et al.*, 2016). Many models project that future climate scenarios could lead to species extinction (Thomas *et al.*, 2004). As the perturbation was carried out in closed bottles, it is easy to exclude the incursion of fresh microbiota. This perturbation, causing a stressful event, might have caused the decline of less adaptive species and the survival of less abundant but more adaptive species. This experiment requires repeating, with additional controls to insure the effects are only caused by the increase in CO_2, This could provide information on the potential effects of climate change on the bacterial community

To conclude, the triplicate independent PCR design herein described was successfully applied to high-depth Illumina single read sequences. Subsampling at various sequencing depths in combination with rarefaction analyses proved the robustness of the proposed method designed. The combination of using PCR replication and singleton removal is therefore proposed as a robust method to define the dominant taxa in any given environment. This was demonstrated by the six distinct habitats, represented by the ten samples analysed, which included both oceanic and costal stations as well as northern and southern hemisphere latitudes. Finally a change in community structure was observed when one of the samples was incubated under future pCO_2 scenario providing a starting point for future experiments on effect of climate change on the bacterial community.

Chapter 3: A full description of the pelagic microbiome (viruses to protists) is possible from a small cup of seawater

3.1 Introduction

Microorganisms dominate the marine environment, reaching 90% of its biomass which can be subdivided into prokaryotes, viruses and protists in increasing order (Suttle, 2007). Although viral biomass count can be estimated to be about 5% of the total biomass, their abundace is proportionally reversed reaching up to 94% of the nucleic acid composition of the oceans (Suttle, 2007). Notwithstanding their abundance very little is known about their diversity in the marine system (Breitbart *et al.*, 2002; Roux *et al.*, 2011), so much so that today we can talk about viruses as the "dark matter" of biology (Pedulla *et al.*, 2003; Roux *et al.*, 2015). It has been estimated that viruses can infect, on a daily basis, a third of the bacterial population (Bergh *et al.*, 1989) and that without the effects of viruses the eukaryotic phytoplankton productivity would increase by 2% (Suttle, 1994). All of these observations show the great importance of studying viruses and their hosts. Nevertheless, the study of marine viruses is complicated by factors such as the lack of conserved genes (Edwards and Rohwer, 2005) as well as difficulties related to laboratory-based cultivation techniques of their hosts (Rappé and Giovannoni, 2003; Edwards and Rohwer, 2005).

In the marine environment the genome size of bacteriophages ranges between 20 and ~250kb (Sandaa, 2008; Steward *et al.*, 2000; Lavigne *et al.*, 2009), whilst for giant viruses genome sizes range from 0.1Mb to 2.5Mb (Colson *et al.*, 2013; Yutin and Koonin, 2013; Claverie *et al.*, 2006; Philippe *et al.*, 2013; Iyer *et al.*, 2001). In recent years the number of studies on marine viruses and their interactions with marine processes has increased, leading to a deeper understanding of this field. Viruses are responsible for significant plankton mortality (Proctor and Fuhrman, 1990; Suttle *et al.*, 1990) which increases overall genetic and biological diversity (Sano *et al.*, 2004). This can be either directly as pathogens causing host mortality (Fuhrman and Noble, 1995; Proctor and Fuhrman, 1990), as well as by restructuring and controlling community composition by a process called “kill the winner” where viruses act as a balancing factor in competing bacterial species (Thingstad, 2000).

Viruses can also influence community structure indirectly through horizontal gene transfer (Sobecky and Hazen, 2009), so much so that mobile genetic elements have been found in marine virus libraries which include hits to bacterial plasmids and various eukaryotic elements (Breitbart *et al.*, 2002). Viruses can, in addition, dramatically change the phenotype of their host via lysogenic conversion (Canchaya *et al.*, 2003). Cell mortality by viral lysis is potentially the most important function of viruses in the aquatic environment, because of its impact on biogeochemical cycles making nutrients more available to small resident microbial communities and cycling carbon faster (Fuhrman, 1992). This viral input plays an important role in the transfer of carbon, nutrients and other elements through the food web and is referred to as the “viral shunt” (Fuhrman, 1999; Wilhelm and Suttle, 1999). The viral shunt favours energy transformation across trophic levels (Roux *et al.*, 2013). A recent study considered the chemical contribution of viral particles to biogeochemical cycles, including supporting phytoplankton growth from the recycling of organically complexed iron (Bonnain *et al.*, 2016).

Despite the increasing awareness of the importance of viruses in key biological processes, major bottlenecks in our understanding of viral diversity and viral roles in marine ecosystems still remain (Roux *et al.*, 2015). Relative to the large diversity of algal species found in the aquatic environment, only a few algal-virus model systems have been studied in any detail. Notable examples of these are the *Emiliania huxleyi* - coccolithovirus (Wilson *et al.*, 2005); ectocarpoids - phaeovirus (Delaroque and Boland, 2008); *Chlorella* - chlorovirus (Yanai-Balser *et al.*, 2010); prymnesiophytes - prasinovirus (Weynberg *et al.*, 2009) and for the photosyntetic bacteria the cyanobacteria - cyanophage (Sullivan *et al.*, 2005) interactions.

The majority, up to 95%, of gene/protein sequences in marine viromes cannot be assigned to known virus genes/proteins or in fact any known entities (Mizuno *et al.*, 2013; Brum *et al.*, 2013b; Angly *et al.*, 2006; Williamson *et al.*, 2012), which causes difficulties in positively identifying viruses within the environment. Despite viruses outnumbering bacteria ranging from 1.4 to 160 (Wigington *et al.*, 2016), this bias is not reflected in the sequences found in metagenomic or genomic databases. Estimates from 2013 based on the European Nucleotide Archive showed that assembled bacterial genomes outnumber marine bacteriophage assembled genomes (3,316 versus 2,010), despite the recent spike in assembled marine phage genomes (Perez Sepulveda *et al.*, 2016). Identification of viruses in the marine environment is made more challenging because some viral genes have been reported to match genes more commonly found in the genomes of their prokaryotic and eukaryotic hosts (Wilson *et al.*, 2005; Baumann *et al.*, 2007; Filée *et al.*, 2007). Therefore, the description of viral diversity has been based on a small and limited number of unique viral genes (Hingamp *et al.*, 2013), often from laboratory cultivated hosts.

The study of viral diversity is complicated further by inconsistencies between methodologies, with processed environmental samples ranging from tens to 400 litres of water (Angly *et al.*, 2006; Venter *et al.*, 2004; Hurwitz and Sullivan, 2013; Williamson *et al.*, 2012). This discrepancy is mainly historical, with sampling of large volumes a necessity for early studies when sequencing technologies required considerable quantities (micrograms) of DNA. In contrast, newer technologies, such as the linear amplification deep sequencing with Illumina, require much smaller quantities (nanograms) of DNA (Hoeijmakers *et al.*, 2011; Marine *et al.*, 2011). Additionally, various sample concentration methods have been developed to collect the greatest quantities of DNA possible from water samples (Lawrence and Steward, 2010; Wommack *et al.*, 2010; John *et al.*, 2011). New methods and technologies present new challenges. Use of standard viral filtration methods involve the use of filters with a pore size of 0.2µm, which removes the bacterial fraction from the sample (e.g. Martínez Martínez *et al.*, 2014), but also leads to the underreporting of the giant virus virions (Claverie *et al.*, 2006; Wilson and Allen, 2009), which can have diameters varying from ~0.2 to 1.5 µm, meaning they will be retained on the filter, with the newly-discovered *Pithovirus sibericum* being the largest member of this group (Legendre *et al.*, 2014, 2015).

The current paradigm of "everything is everywhere" (Angly *et al.*, 2006; Breitbart and Rohwer, 2005) suggests that all the major virus taxa can be found everywhere. This is largely due to the presence of cyanophage-like sequences, of the order *Caudovirales*, dominating all oceans' viromes (Angly *et al.*, 2006; Munn, 2006; Breitbart *et al.*, 2002), including the recently sampled Indian Ocean (Williamson *et al.*, 2012). The order *Caudovirales* comprises three families: *Myoviridae* (contractile tails), *Siphoviridae* (non-contractile tails) and *Podoviridae* (short tails) (Ackermann, 2003). During the Global Ocean Sampling expedition (GOS) (Williamson *et al.*, 2008), myovirus-related sequences were ubiquitously distributed among sampling sites, with the highest prevalence at tropical oligotrophic locations. Podo- and siphoviruses showed site-specific distributions, with their highest abundance in temperate mesotrophic waters and hypersaline lagoons respectively (Williamson *et al.*, 2008). Within the Indian Ocean 32% of the viral fraction (VF) was attributed to known viruses, with 95% of the known viruses identified as belonging to the order *Caudovirales* (*Myoviridae* 54.3%, *Podoviridae* 27.6%, *Siphoviridae* 17%) (Williamson *et al.*, 2012). The NCLDVs were often the next major lineage present, with the family *Phycodnaviridae* representing 83.9% of this group, followed by presumptive members of the *Iridoviridae* (8.5%) and *Mimiviridae* (7.3%) families (Williamson *et al.*, 2012).

To date, most viromic studies have not reported on the diversity of the potential hosts that the viruses infect, making it unclear as to whether the viruses present in the water column are the result of active or past infections. In contrast, the Tara Oceans expedition reported on the eukaryotic and prokaryotic diversity (de Vargas *et al.*, 2015; Sunagawa *et al.*, 2015) in conjunction with the viral diversity (Brum *et al.*, 2015a; Mihara *et al.*, 2016). Global prokaryotic diversity has shown high abundance of Alphaproteobacteria in surface waters (SRF) and at the deep chlorophyll maximum (DCM). The second most represented group is the Cyanobacteria and Gammaproteobacteria, at varying proportions depending on locations. An exception to these results was found in the south-west Indian Ocean, which was dominated by Cyanobacteria taxa, then Gammaproteobacteria and finally Alphaproteobacteria (Sunagawa *et al.*, 2015).

For the eukaryotic fraction, samples collected during the Tara Ocean expedition showed that the pico-nanoplankton fraction was dominated by photosynthetic dinoflagellates, class Dinophyceae. However, the heterotrophic protists showed the highest richness and abundance across all the other size fractions. Parasites of the order Alveolata, known to routinely infect the Dinophyceae, mainly consisted of members of the order Syndiniales, specifically the MALV- I and MALV-II clusters (new nomenclature Syndiniales groups I and II, Horiguchi, 2015), up to 88% of abundance across some stations. For the south-west Indian Ocean, the eukaryotic fraction was dominated by alveolates including the Dinophyceae, and parasites such as MALV and Syndiniales taxa (http://taraoceans.sb-roscoff.fr/EukDiv/static/files/krona/krona.TV9_52.html) (de Vargas *et al.*, 2015).

In this study, the hypothesis that a volume equivalent to a cup of seawater (250 ml) is sufficient to describe the most abundant microbial taxa (from viruses to protists) in the marine environment, will be tested. Serendipitously, our study site is within 548 nautical miles of station 64 previously sampled by the Tara Oceans expedition (-29.5333, 37.9117), thereby allowing for a semi-qualitative comparison to be made. Our protocol differed from previous studies, including that of Tara Oceans, as it contained no concentration steps after water was collected. In addition, only 50ml of the 0.45µm 250ml permeate was used to describe the combined virus fraction, small bacteria (Tabor *et al.*, 1981), vesicle (Biller *et al.*, 2016) and free DNA fractions (eDNA) (Taberlet *et al.*, 2012). The 0.45µm size fraction was chosen in order to limit the removal of the giant viruses. Here it will be reported that a relatively small water sample can be used to capture the dominant microbial taxa within any given aquatic system. Moreover, microbial diversity can now be assessed alongside the traditional oceanographic conductivity, temperature and depth (CTD) measurements taken from the identical water sample collected from the same body of water.

3.2 Materials and Methods

3.2.1 Sample collection

The water sample was collected during the second transect of the Great Southern Coccolithophore Belt expedition (GSCB-cruise RR1202) in the south-west Indian Ocean in February 2012 (http://www.bco-dmo.org/project/473206). The location of the sampling station S1 (-38.314983, 40.958083, water temperature 20.83°C, pH 8.08) was mapped using RgoogleMaps_1.2.0.7 (Loecher and Ropkins, 2015) under R version 3.3.0 (2016-05-03) (Figure 3.1).

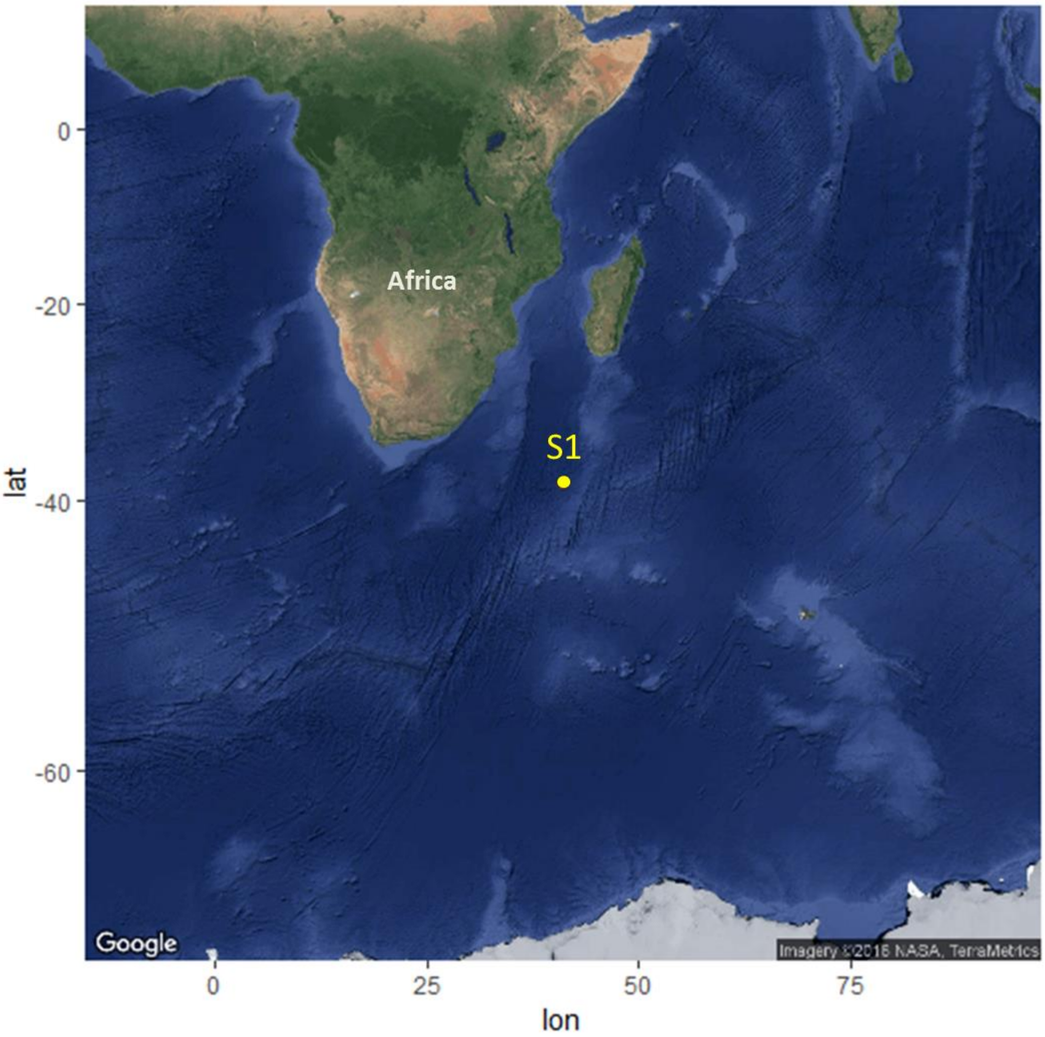


Figure 3.1: Latitude (-38.314983) and longitude (40.958083) of sample station S1.

One litre of water was gathered from each conductivity-temperature-depth (CTD) cast at the chlorophyll maximum (5m). From this, an aliquot of 250ml of seawater was filtered through a 0.45µm polycarbonate filter and the filter was used for the DNA extraction on-board the R/V Roger Revelle, using Qiagen DNeasy Blood and Tissue protocol (QIAGEN, Valencia, CA, USA). The DNA was stored at -20°C and subsequently transferred to Plymouth, UK, for further processing. Fifty ml of filtered water were set aside, wrapped in tin foil stored in a fridge in the dark. This too was returned to Plymouth, UK, for further processing.

3.2.2 DNA extraction, preparation and sequencing of the >0.45 µm fraction

DNA was extracted following the protocol described in Chapter 2. The V4 region along the prokaryotic 16S ribosomal RNA gene (rRNA) was amplified using the universal primer pair 515F and Illumina tagged primer 806R7, 806R10 and 806R15 (Caporaso *et al.,* 2012a, 2011). For the eukaryotic 18S ribosomal RNA gene, the primer pair 1391F and Illumina tagged EukB6, EukB16 and EukB23 was used to amplify the V9 region (Stoeck *et al.*, 2010). PCR reactions contained 10ng of environmental DNA, to 5X Colourless GoTaq Flexi Buffer (Promega) 1 µl of Forward and Reverse Primers (10pmol) (Appendix I), 1.5µl $MgCl_2$ Solution 25mM (Promega), 2.5µl dNTPs (10mM final concentration, Promega), 1µl EvaGreen Dye 20X (Biotium), 0.1µl GoTaq DNA Polymerase (5u/µl- Promega) and made up to a final volume of 25µl with sterile water for each reaction (Table 3.1). This was done to determine the mid-exponential threshold of each reaction, which were run on a Corbett Rotor-Gene™ 6000 (QIAGEN, Valencia, CA, USA). The real time PCR comprised of an initial denaturation at 94 °C for 3 minutes, followed by 40 cycles of a three step PCR: the cycles were 94°C for 45 seconds, 50°C for 60 seconds and 72°C for 90 seconds. The fluorescence was acquired at the end of each annealing/extension step on the green channel. The cycle threshold of the amplification in the exponential phase was recorded for amplification.

A second standard PCR amplification was carried out in triplicate and run under the same conditions, excluding the addition of the Evagreen Dye. The sample was removed from the machine when it reached the cycle threshold as previously determined to prevent PCR bias. Products were run on a 1.4% agarose gel to confirm the success of the amplification and the product size of the amplification. The bands were cut out and purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research). Quantity and quality was verified on the NanoDrop 1000 (Thermo Scientific) and QuantiFluor E6090 (Promega). V4-16S and V9-18S were prepared mixing an equimolar concentration of each amplicon triplicate into the pool for which concentration was checked on the Bioanalyser (Agilent). The final pooled samples were denatured and diluted to 6pM and mixed with 1pM PhiX control (Illumina), forwar read (read 1) sequencing primer was diluted in HT1, before the flowcell was clustered on the cBOT (Illumina). Multiplexing sequencing primers and reverse read (read 2) sequencing primers were mixed with Illumina HP8 and HP7 sequencing primers, respectively. The flowcell was sequenced (100 paired end) on HiSeq 2000 using SBS reagents v3. The raw sequences have been deposited at the European Nucleotide Archive (ENA) under accession number PRJEB16346.

Table 3.1: PCR protocol, reaction setup

Reagent	Volume
Forward Primer (10pmol)	1µl
Reverse Primer (10pmol)	1µl
5X Colourless GoTaq Flexi Buffer	5µl
dNTP (10mM)	2.5µl
$MgCl_2$ Solution 25mM	1.5µl
GoTaq Polymerase (5u/ µl)	0.1µl
EvaGreen Dye 20X	1µl
DNA	1 to 5µl to give a concentration of 10 ng
Sterile water	To find a total reaction volume of 25µl

3.2.3 DNA extraction, preparation and sequencing of the <0.45 µm fraction

The whole 50ml permeate was used in the nucleic acid extraction procedure. To the permeate was added 100µl of proteinase K (10mg/ml; Sigma-Aldrich) and 200µl of 10% SDS (Sigma-Aldrich), the solution was then incubated for 2 hours with constant rotation at 55°C. The lysate was collected by multiple centrifugations on a Qiagen DNeasy Blood and Tissue column (QIAGEN, Valencia, CA, USA). The standard Qiagen protocol was followed with 20µl nuclease-free water (SIGMA) used as the elution agent. Quantity and quality was determined using the NanoDrop 1000 (Thermo Scientific) and QuantiFluor E6090 (Promega). 200µl DNA

(< 40ng) was fragmented by sonication using a Bioruptor (Diagenode) on medium for 15 bursts of 30s, with a 30s pause the resulting solution was then concentrated to 30µl on a MinElute column (QIAgen). Fragments were made into libraries using the Nextflex ChipSeq library preparation kit (Bioo Scientific) without size selection and with 18 cycles of PCR amplification as part as library enrichment, Nextflex adapter sequences are illustrated on Appendix III. . Bioanalyser (Agilent) analysis indicated the final library contained inserts between 30bp to 870bp. The library was multiplexed with other samples and sequenced (100 paired end) on a HiSeq 2000 (Illumina) using RTA1.9 and CASAVA1.8. The raw sequences have been deposited at the European Nucleotide Archive (ENA) under accession number PRJEB166674.

3.2.4 Bioinformatic pipeline for the prokaryotic (16S) and eukaryotic (18S) amplicon

The complete bioinformatic pipeline is illustrated in Figure 3.2. Analyses of the amplicon datasets (16S and 18S rRNA) were performed using the Bio-Linux 8 system at the Marine Biological Association of the UK, whilst computations of the metagenome dataset were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: http://hpc.uct.ac.za.

The read quality was first assessed using Fast-QC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) was utilised for the trimming and filtering steps; the first and last 10 bases were trimmed in order to remove low quality nucleotides. Reads were then filtered in order to retain only reads with more than 95% of nucleotide positions called with a quality score of 20. Trimmed and cleaned reads (Table 3.2) from each of the triplicate V4-16S and V9-18S PCRs were pooled in order to assign Operational Taxonomic Units (OTUs) using Qiime (Caporaso *et al.*, 2010) with 97% similarities for clustering and Swarm analysis (Mahé *et al.*, 2014), respectively. Taxonomy was assigned using BLASTn implemented in Qiime and Swarm using SILVA v119 (https://www.arb-silva.de) with a minimum e-value of 10e-05.

Table 3.2: Description of sequences. (*) Indicates number of OTU and phylotypes when the T0 and T1 threshold were applied to a combination of the three replicates, duplicate values have been removed.

Sample	Processed reads	Sequence length	OTU T0	Phylotypes T0	OTU T0*	Phylotypes T0*	OTU T1	Phylotypes T1	OTU T1*	Phylotypes T1*
16S Rep1	741,033	125	20,381	882	45,826	1,409	11,341	561	23,081	834
16S Rep2	1,117,576	125	30,642	1,077			16,593	697		
16S Rep3	841,639	125	24,756	767			13,416	505		
18S Rep1	223,814	125	2,972	339	6,836	477	1,714	267	2,930	346
18S Rep2	275,201	125	3,271	353			1,780	279		
18S Rep3	308,208	125	3,470	346			1,836	278		
Metagenome	10,036,627 (bp)	Average 78.9								
Contigs (4,962)		min: 240 max: 74,442 Average: 1,045		254 (virus)						

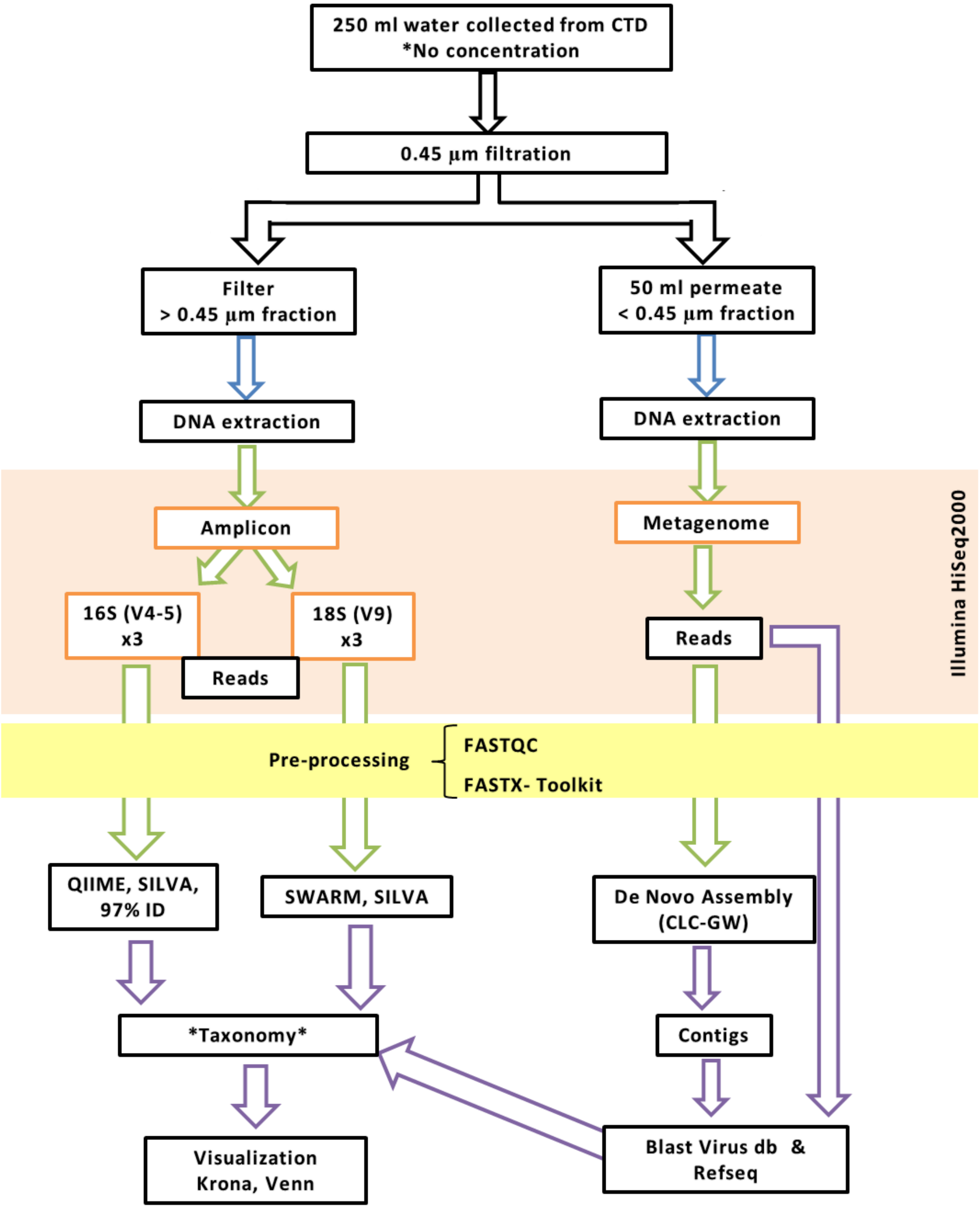


Figure 3.2: Schematics of the bioinformatic pipeline.

3.2.5 Bioinformatics pipeline of the 0.45µm permeate

As was previously done for the amplicon dataset (Chapter 2), the quality of the reads was first assessed using Fast-QC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) was used to trim the first last bases to remove low quality nucleotides, and subsequently to filter out reads with fewer than 95% of nucleotide positions called with a quality score of 20. The forward read (read 1) of the 100 bp pair-end HiSeq reads were subjected to random library size normalization using Qiime script subsample_fasta.py (Caporaso *et al.*, 2010); reverse read (read 2) had poor quality and was therefore discarded. The reads were used in a BLASTx (Altschul *et al.*, 1990) analysis against a Virus database (db; courtesy of Pascal Hingamp), containing Refseq curated viral genomes together with additional new genomes (Mihara *et al.*, 2016). 20% of R1 Refseq whole organism db (Tatusova *et al.*, 2014) was used as reference database for the analyses with an e-value less than 10e-05. BLAST analyses were performed on the University of Cape Town's hex cluster. In addition, the pair-end reads were then assembled into contigs using a De-Bruijin *de novo* assembly program in CLC Genomic Workbench version 7.1.5 (CLCbio, Cambridge, MA, USA) using global alignment with automatics bubble and word size, minimum contigs length of 250, mismatch cost of 2, insertion and deletion cost of 3, length fraction of 0.5 and similarity threshold of 0.8 (Table 3.1). The contigs were estimated with the BLASTx as described for the R1 normalised reads.

The top hits from all the blast searches were selected through the use of a parser Perl script (http://www.bioinformatics-made-simple.com). The ICTV database 2013 v1 implemented with the NCBI taxonomy were utilised to create a viral taxonomy catalogue, this was then merged, using R, with the BLAST output to assign taxonomy. Affinity of sequences with the order *Megavirales* was assigned according to recent publications (Koonin and Yutin, 2010; Colson *et al.*, 2012; Filée, 2013).

3.2.6 Visualization of community diversity

Krona tools (Ondov *et al.,* 2011) were used to visualize community diversity as evaluated by the Silva (v119), Refseq and Virus db genes taxonomy assignments. Venn diagrams were created using the R package VennDiagram_1.6.17 on R version 3.3.0 (2016-05-03).

3.2.7 Thresholds applied to annotated datasets

Based on the analyses carried out for Chapter 2, independent analyses were performed on the three replicates (V4-16S and V9-18S) and assigned taxonomy using Silva (Pruesse *et al.,* 2007). The use of replication was aimed to the removal of noise in the sample while keeping the rare organisms as shown in Chapter 2. Modification of level of stringency (filters) applied in the previous chapter will be therefore considered: (1) T0, all phylotypes present across the three replicates; (2) T1, removing singletons from each replicate; (3) T10, a minimum of 10 copies per phylotype had to be present in any one of the replicates, (4) T10-R1, a minimum of 10 copies per phylotype present in any two replicates and (5) T10-R2, a minimum of 10 copies per phylotype present in all three replicates.

3.3 Results

3.3.1 Microbiota captured on the 0.45 µm filter

After pre-processing, which included a specific sub-sampling to an equal read length of 125 bases, on average 900,082 reads were retained for the prokaryotic and 269,074 for the eukaryotic dataset (Table 3.2). These reads clustered (T0 applied to combination of the three replicates) into approximately 46 thousand unique OTUs for the prokaryotes, consisting of which clustered into 1,409 phylotype (mostly at the taxonomic level of species). For the eukaryotes 6,836 OTUs clustered into 477 phylotypes (Table 3.2). The application of four thresholds, to implement the work carried out in chapter 2, resulted in an increase in selection stringency (T0 to T10-R2) without the removal of significant numbers of reads from the prokaryotes (Figure 3.3a) and eukaryotes (Figure 3.4a) datasets, independent of sequence depth. However, the greatest change observed due to the application of the thresholds, was seen in the number of phylotypes observed (Figures 3.3b and 3.4b). A total number of 1,886 phylotypes were observed in the 250ml sample from the south-west Indian Ocean, made up of 1,409 prokaryotic and 477 eukaryotic phylotypes. When the singletons were removed (T1), the number of prokaryotic phylotypes dropped by nearly a half to 834 (59.19%, phylotypes retained) (Figure 3.3b); this was also observable in the OTUs (Table 3.2) moving from 45,826 to 23,081. Similarly, the number of eukaryotic phylotypes dropped by a third to 346 phylotypes (72.54% phylotypes retained) (Figure 3.4b), whilst OTUs dropped from 6,836 to 2,930 (Table 3.2). When a further threshold T10 was applied (i.e. the criteria that there must be a minimum of 10 reads per phylotype in any of the replicates), the diversity dropped from T0 by an additional 36% to just under 77% for prokaryotes - retaining only 23% (Figure 3.3b), and 24% to 51% in eukaryotes – retaining only 49% (Figure 3.4b), leaving a total number of phylotypes as 554.

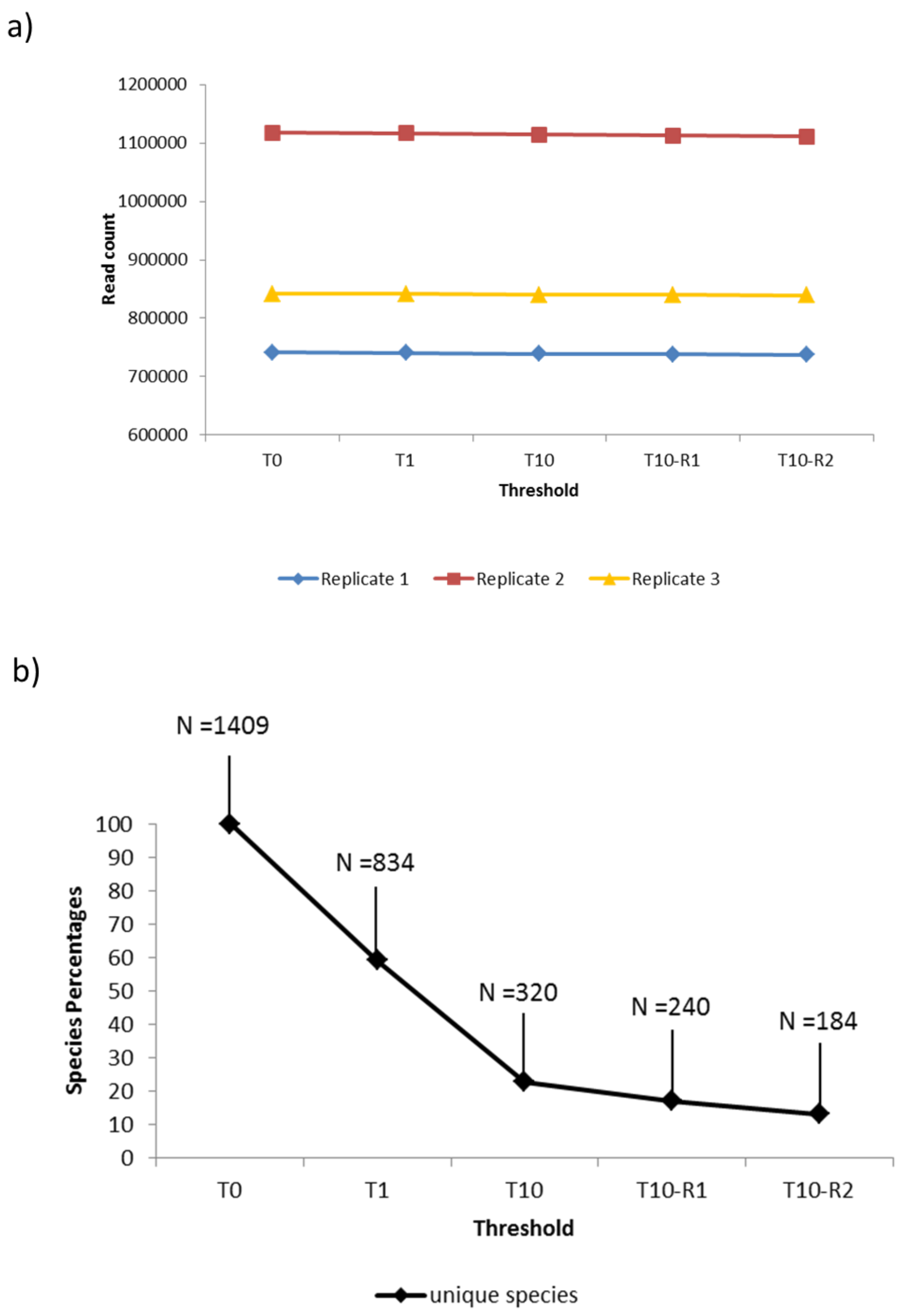


Figure 3.3: Analysis of the prokaryotic fraction. a) Reduction in the number of reads when thresholds are applied. b) Percentage of reads and phylotypes counted when thresholds are applied

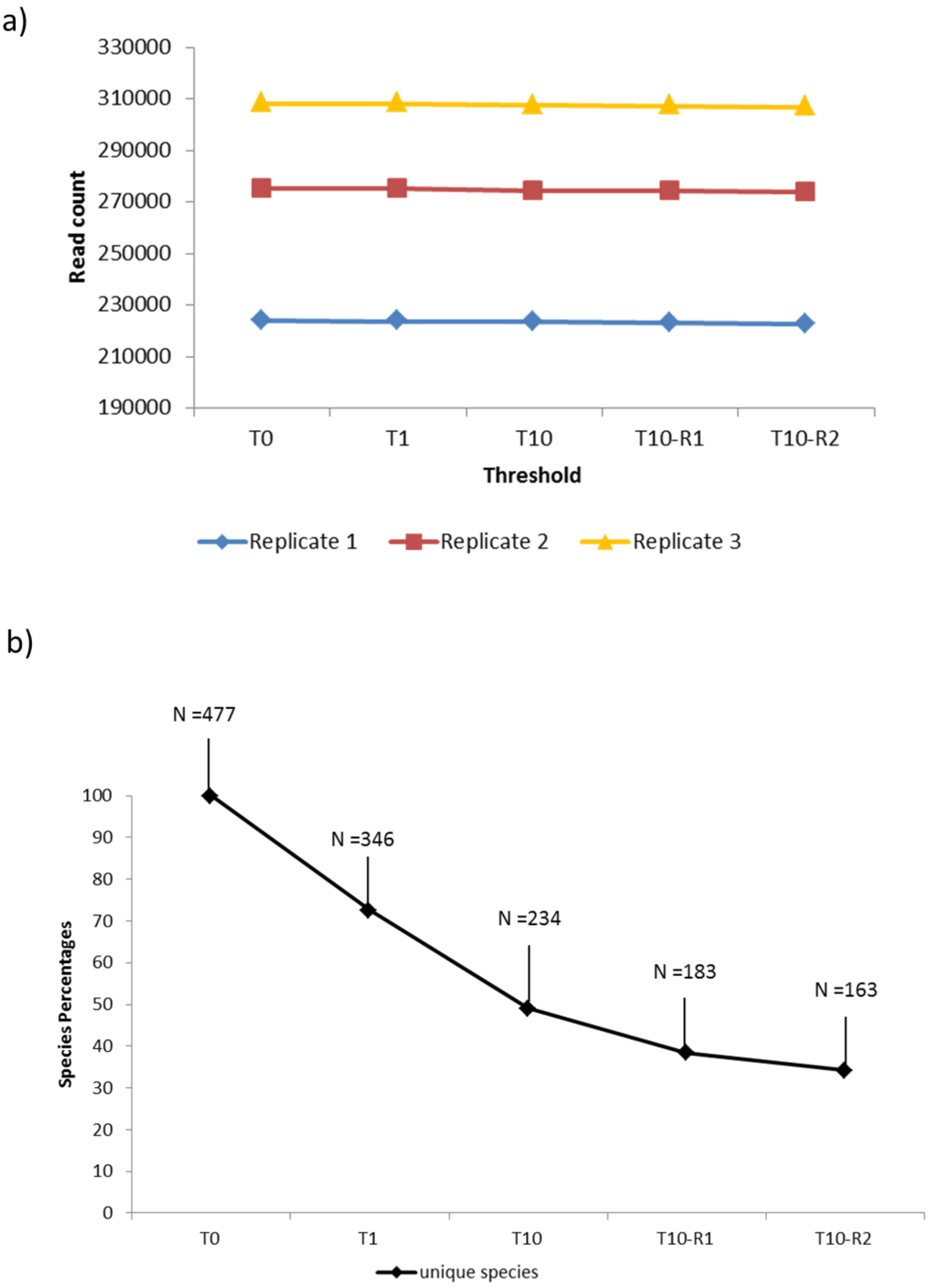


Figure 3.4: Analysis of the eukaryotic fraction. a) Reduction in the number of reads when thresholds are applied. b) Percentage of reads and phylotypes count when thresholds are applied

The phylotypes removed after applying the singleton threshold (T1) (Supplementary Table 4) included *Cicer arietinum* (chickpeas), *Sesamum indicum* (sesame) and *Nicotiana sylvestris* (tobacco), which were not expected to be present in the marine environment. The application of the T10 threshold resulted in the removal of a few marine species instead, such as *Noctiluca scintillans*, *Amphidinium mootonorum* and *Pandorina morum*. The additional application of replication thresholds, present in greater than ten copies in at least any two (T10-R1) and all three (T10-R2) replicates, revealed a further but minimal reduction in the overall phylotype content (Figure 3.3b & 3.4b): both the prokaryotes and eukaryotes dropped to 17% and 38% (from T10 to T10-R1, Figure 3.3b) and 13% and 34% (from T10 to T10-R2, Figure 3.4b), respectively. A core of 184 phylotypes could be identified for the prokaryotes (Figure 3.5a) and 163 for the eukaryotes (Figure 3.5b), which were common across all thresholds. If no threshold was applied, 575 prokaryotes (41%) and 131 eukaryotes (27%) unique or rare were observed, however, irrespective of which threshold is applied no phylotype unique to their stringency were observed (Figures 3.5).

In summary, a total of 1,886 phylotypes have been identified without the application of any threshold (T0), which was reduced to 1,180 after singletons has been removed (T1). A further decrease in phylotype composition to 554, 423 and 347 has been identified after application of T10, T10-R1 and T10-R2 thresholds.

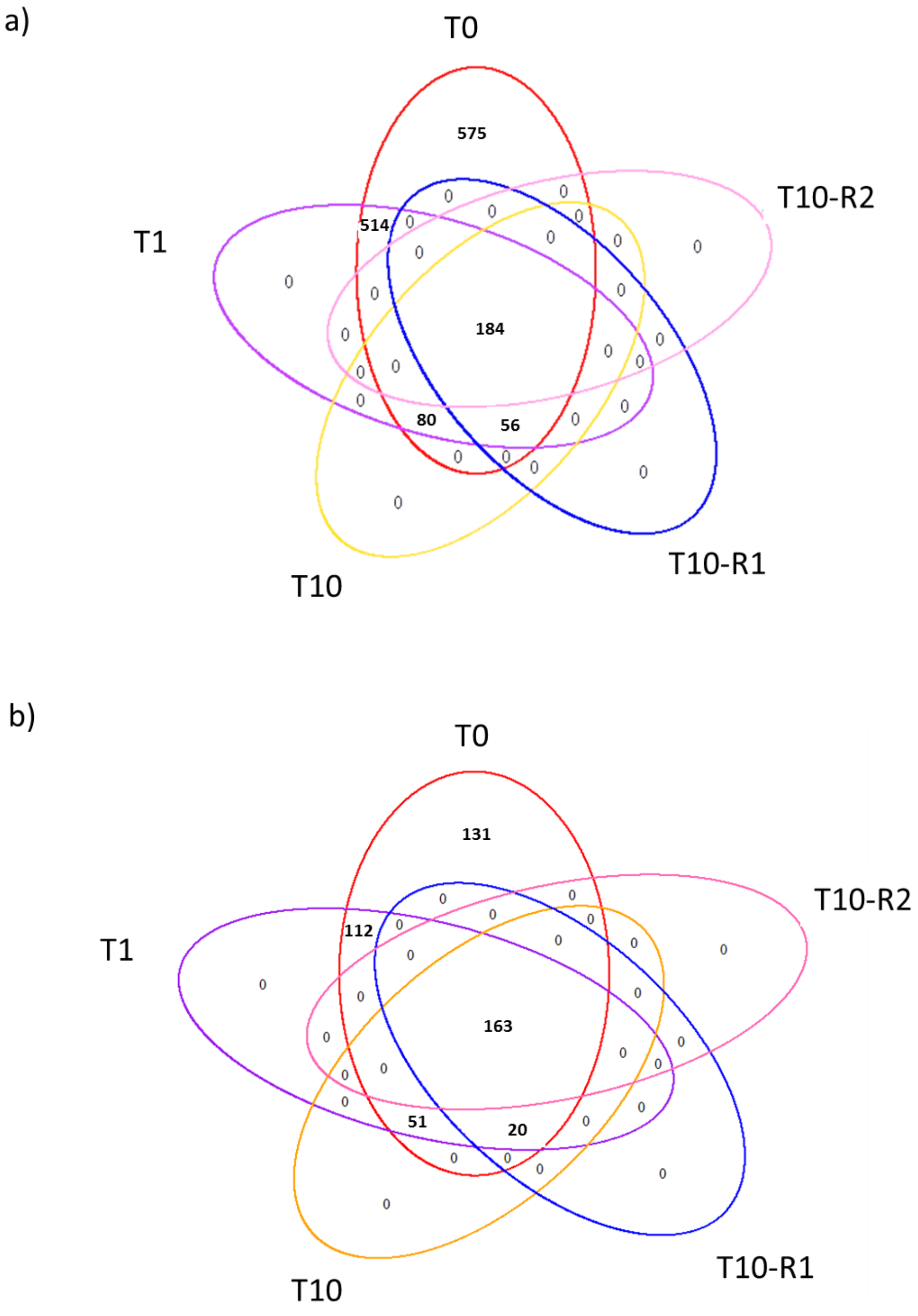


Figure 3.5: Presence absence analyses at phylotypes level before and after the application of the thresholds. a) prokaryotes (16S), b) eukaryotes (18S).

The three replicates have been considered independently in order to understand how phylotypes differ across the three PCR replicates (Figure 3.6). Prokaryotic diversity ranged from 767 phylotypes in replicate 3 to 1,077 in replicate 2 (Figure 3.6a), corresponding to the sequence depth (Figure 3.3a). This was not observed for the eukaryotes (Figure 3.6b), which ranged from 339 of replicate 1 to 353 of replicate 2 (Figure 3.4d), irrespective of the sequence depth (Figure 3.4a). When applying the T1 threshold, the number of phylotypes retained were on average 65% (from 882 to 561 in replicate 1, from 1,077 to 697 replicate 2 and from 767 to 505 in replicate 3) and 79% (from 339 to 267 in replicate 1, from 353 to 279 in replicate 2 and from 346 to 278 in replicate 3) of the prokaryotes and eukaryotes, respectively (Figure 3.6). Applying stringency threshold T10 reduced the prokaryotic diversity in replicate 1 to 28%, in replicate 2 to 27% and replicate 3 to 26% (Figure 3.6a), whilst for the eukaryotes across replicates 1, 2 and 3 to 57%, 55% and 58% respectively (Figures 3.6b).

Phylotype composition at T0 had 36% prokaryotic and 50% eukaryotic phylotypes in common across all replicates (Figures 3.7). Between 9 and 22 % percent of prokaryotes and 10 and 22% of eukaryotes were unique to each replicate. When singletons (T1) were removed and the T10 threshold applied, the phylotypes common across all replicates increased to 45% and 58% for prokaryotes (Figure 3.7a), whilst for the eukaryotes, increased to 61% and 70% (Figure 3.7b). This coincided with the reduction in unique phylotypes retained per replicate. Replicate 1, 2 and 3 changing from 164 to 22, 309 to 55 and 124 to 2 unique prokaryotic phylotypes (Figure 3.7a). Similarly, replicate 1, 2 and 3 changed from 48 to 16, 57 to 12 and 49 to 16 unique eukaryotic phylotypes (Figure 3.7b).

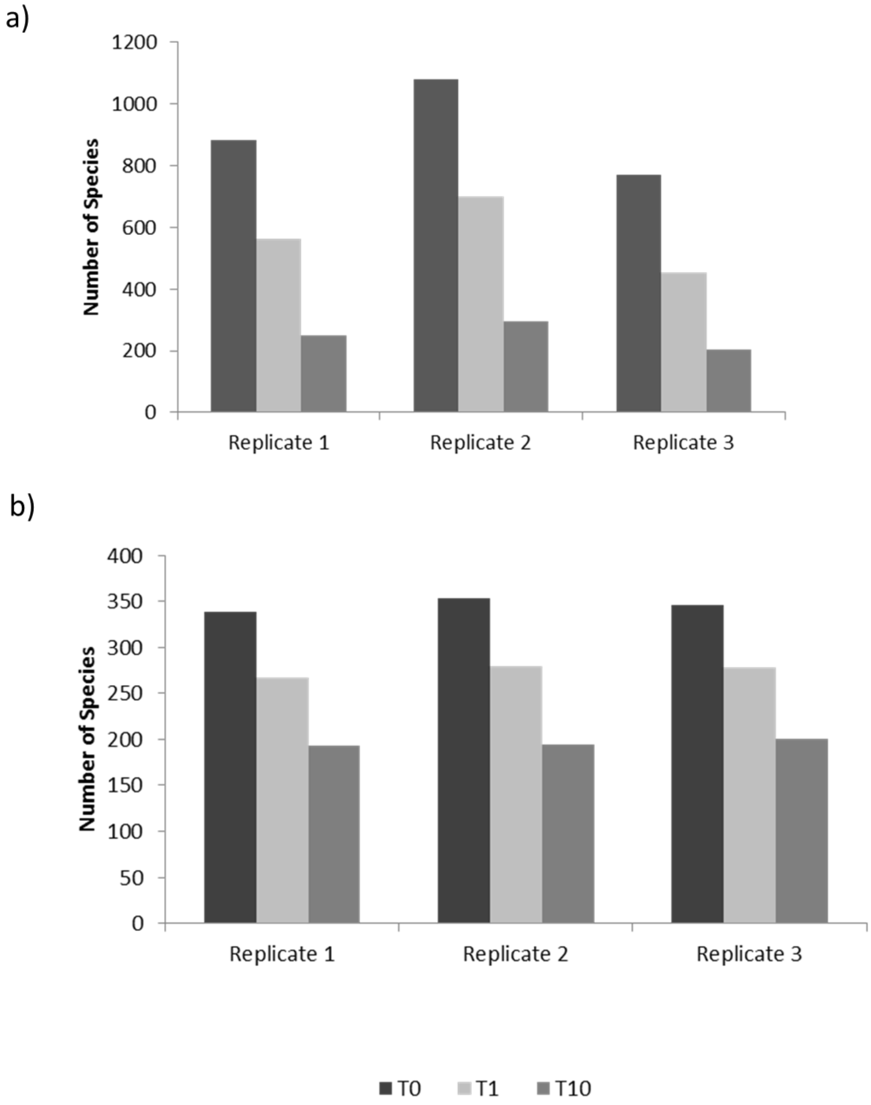


Figure 3.6: Analysis of phylotypes by replicate for the prokaryotes (a) and eukaryotes (b).

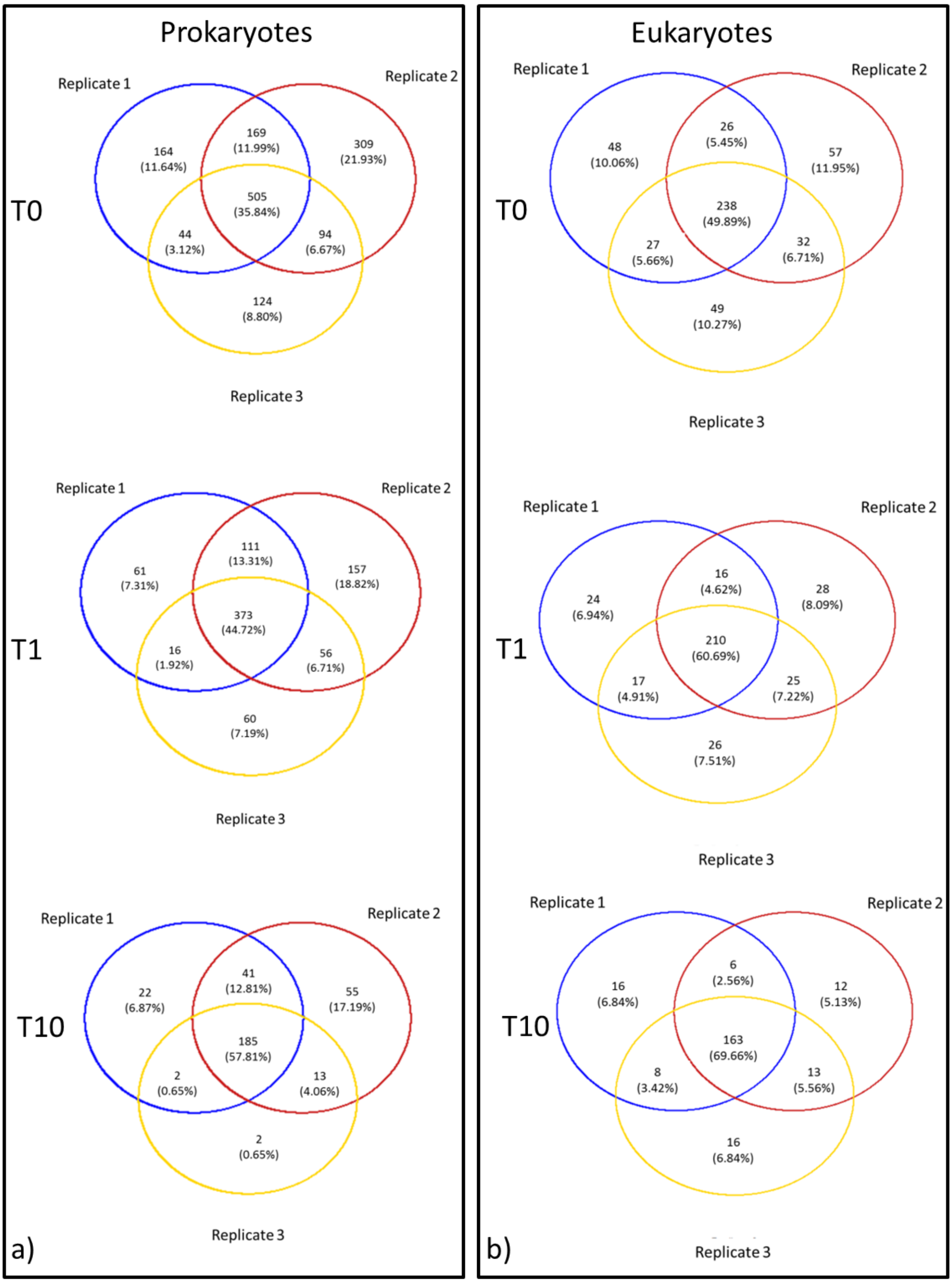


Figure 3.7: Presence absence analysis at phylotypes level when thresholds are applied to each individual replicate.

3.3.2 Diversity and community structure of the >0.45 µm fraction (T1)

Cyanobacteria made up 41.88% of the prokaryotic community diversity; its composition was dominated by members of the genera *Synechococcus* (30%) and *Prochlorococcus* (9%) (Table 3.3, Supplementary Figure 1). The V4-16S universal primers also amplified the eukaryote plastid ribosomal genes, making up 2.68% of the total sequences (Table 3.2, Supplementary Figure 1). The second main bacterial group was Proteobacteria (32.14%) comprising the classes Alphaproteobacteria (19.75%), Gammaproteobacteria (8.17%) and Deltaproteobacteria (3.16%). The Alphaproteobacteria class can be further separated into the orders Rhodospirallales (5.05%), Rickettsiales (4.80%), Rhodobacteriales (4.49%) and the clade SAR11 clade (4.93%). Gammaproteobacteria's was comprised by the orders Oceanospirallales (5.57%), Alteromonadales (0.84%) and Marinicella (0.74%). The class Deltaproteobacteria consisted mainly of the SAR324 clade (2.85%). Bacteroidetes and Actinobacteria represented 3.26% and 1.67% of the prokaryote diversity (Table 3.3, Supplementary Figure 1). Finally, a large component of the prokaryotic community could not be assigned (20%).

The eukaryotic community was dominated by the group Alveolata (91.66% of the eukaryotes) (Table 3.3, Supplementary Figure 2), comprising the Protoalveolata (43.86%) and Dinoflagellata (41.35%), Ciliophora (3.40%) and FV18-2D11 (2.70%). Protoalveolata were characterised for 97% (42% of the total sequences) by Syndiniales subdivided as: Group II (57%), Group I (18%), Amoebophyra (17%) and Duboscquella (4%) and Perkinsidae (3%). The group Dinoflagellata was formed by Peridiniphycidae (16%), Gymnodiniphycidae (14%), Dinophysiales (1%) and Prorocentrum (0.7%).

Table 3.3: Composition of prokaryotic and eukaryotic (>0.45µm fraction) using Silva database (v 119). Contigs (<0.45 µm fraction, permeate) annotation using Refseq database. T1 average for three replicates shown. Shown only values >0.5%

Dataset	L1	L2	L3	
Prokaryotes	**Bacteria (80.11%)**	**Cyanobacteria (41.88%)**	Cyanobacteria (39.14%)	*Synechoccoccus* (30%) *Prochlorococcus* (9%)
			Chloroplast (2.68%)	
		Proteobacteria (32.14%)	Alphaproteobacteria (19.75%)	Rhodospirillales SAR11 clade (4.93%) Rickettsiales (4.80%) Rhodobacterales
			Deltaproteobacteria	SAR324 clade (2.85%)
			Elev-16S-509 (0.87%)	
			Gammaproteobacteria (8.17%)	Alteromonadales Oceanospirillales *Marinicella* (0.74%)
		Bacteroidetes (3.26%)	Flavobacteriia (3.12%)	Flavobacteriaceae NS9 marine group
		Actinobacteria	Acidomicrobiia	OM1 clade (1.55%)
		Verrucomicrobia (0.5%)		
	No blast hit (20%)			
Eukaryotes	**Eukaryota (99.72%)**	**Archaeplastida (2.68%)**	Chloroplastida	Chlorophyta (1.88%)
			Rhodophyceae	Florideophycidae
		Cryptophyceae	Cryptomonadales (0.67%)	
		Haptophyta (2.33%)	Pavlovophyceae	*Diacronema* (1.09%)
			Prymnesiophyceae	Prymnesiales (0.65%)
		SAR (92.92%)	Alveolata (91.66%)	Ciliophora (3.40%) Dinoflagellata (41.35%) FV18-2D11 (2.70%) Protoalvolata (43.86%)
			Rhizaria (1.14%)	Retaria (1.04%)
		uncultured marine eukaryote (0.77%)		
Metagenome RefSeq	**Bacteria (86.85%)**	**Actinobacteria (47.30%)**	Actinobacteria (47.15%)	Corynebacteriales Micrococcales (40.69%) Propionibacteriales Pseudonocardiales Streptomycetales
		Firmicutes (0.67%)		
		Proteobacteria (38.20%)	Alphaproteobacteria (36.51%)	Rhizobiales (0.89%) Rhodobacterales Sphingomonadales
			Betaproteobacteria (0.56%)	
			Gammaproteobacteria (1.05%)	
	Eukaryota	**Phaeophyceae**	Ectocarpales (0.79%)	
	Virus (0.75%)			
	No blast hit (11.03%)			

3.3.3 Diversity in <0.45µm fraction

After pre-processing, 10 million paired reads were utilised for contigs assembly with an average contig length of 1,045 bp (Table 3.2), whilst a sub-sampled of 1.5 million reads from the forward read were utilised for analysis at reads level. The majority of sequences and predicted genes based on blastx against a virus database could be annotated as "other than virus" (Figure 3.8a & 3.8b). This was independent of whether the reads (99%, Figure 3.8a) or the assembled contigs (86%, Figure 3.8b) were used for the annotation. Using the Refseq database, the metagenome could be divided into 59.92% Bacteria, 39.32% unknown, 0.71% Eukaryota and 0.05% viruses at the reads level, whilst for the contigs the hits could be divided into Bacteria (86.85%), unknown (11.03%), Eukaryota (1.35%), viruses (0.75%) and Archaea (0.02%) (Figure 3.8c & 3.8d).

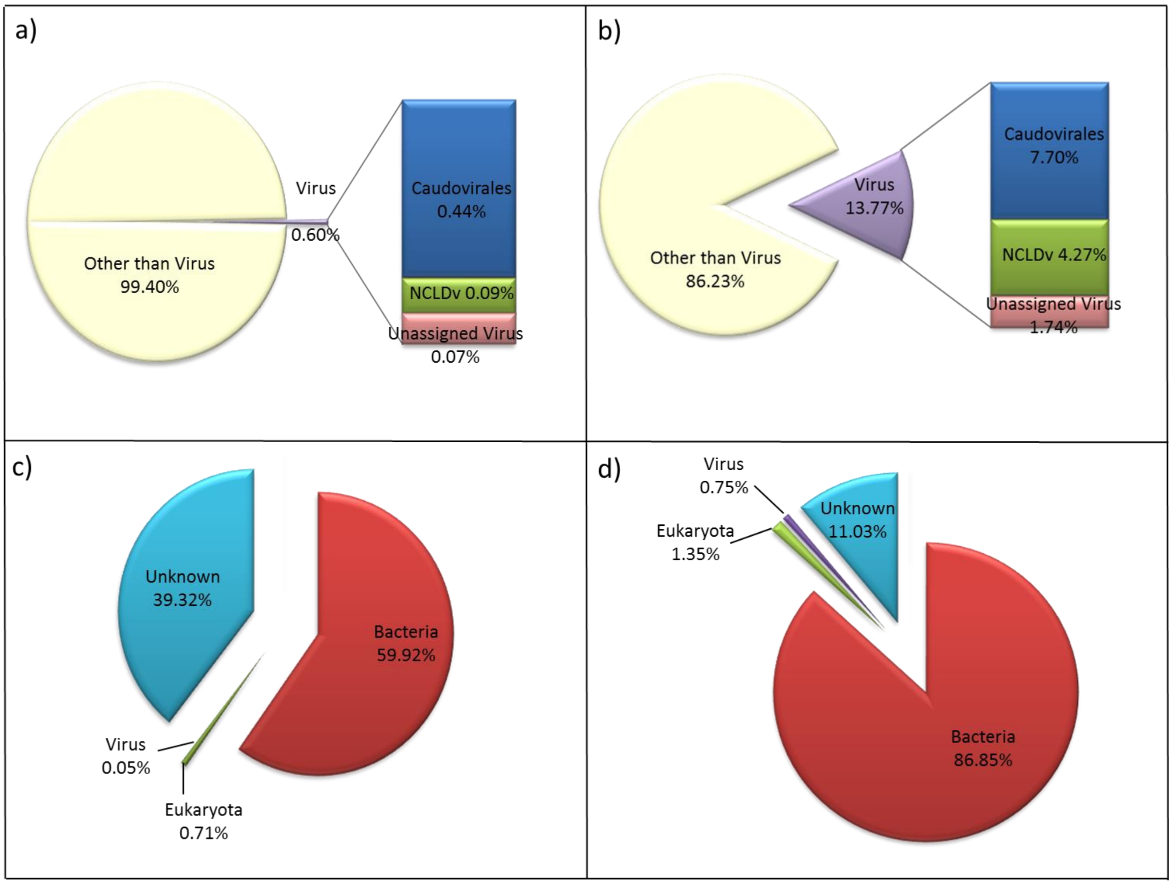


Figure 3.8: Taxonomic assignment based on reads (a, c) and contigs (b, d). Reads (forward read only) were annotated using (a) the Virus database and (c) the Refseq database; contigs were annotated using (b) the virus database and (d) the Refseq database.

Utilising the output from the Refseq database, the annotation for reads and contigs were compared. It is possible to observe very low similarities between the phylotypes annotated in the reads versus the contigs (Figure 3.9, Supplementary Figure 3). Only 8.81% of phylotypes were common across the two methods when no threshold was applied (T0) (Figure 3.9a), whilst 13.35% were common when T10 was applied (Figure 3.9c). To account for the high level of randomness associated with the top hits from BLAST outputs, especially from universal conserved genes, the analysis were repeated using a lower stringency annotation, i.e. the genus as lowest level of classification instead of the phylotypes. Common annotations between the analysis, based on reads versus contigs, increased to 17.93% at T0 (Figure 3.9b) and 37.48%

at T10 (Figure 3.9d). Therefore, due to the major data loss that will be encountered if reads were utilised, from here on, only the annotation based on the contigs will be the method of choice.

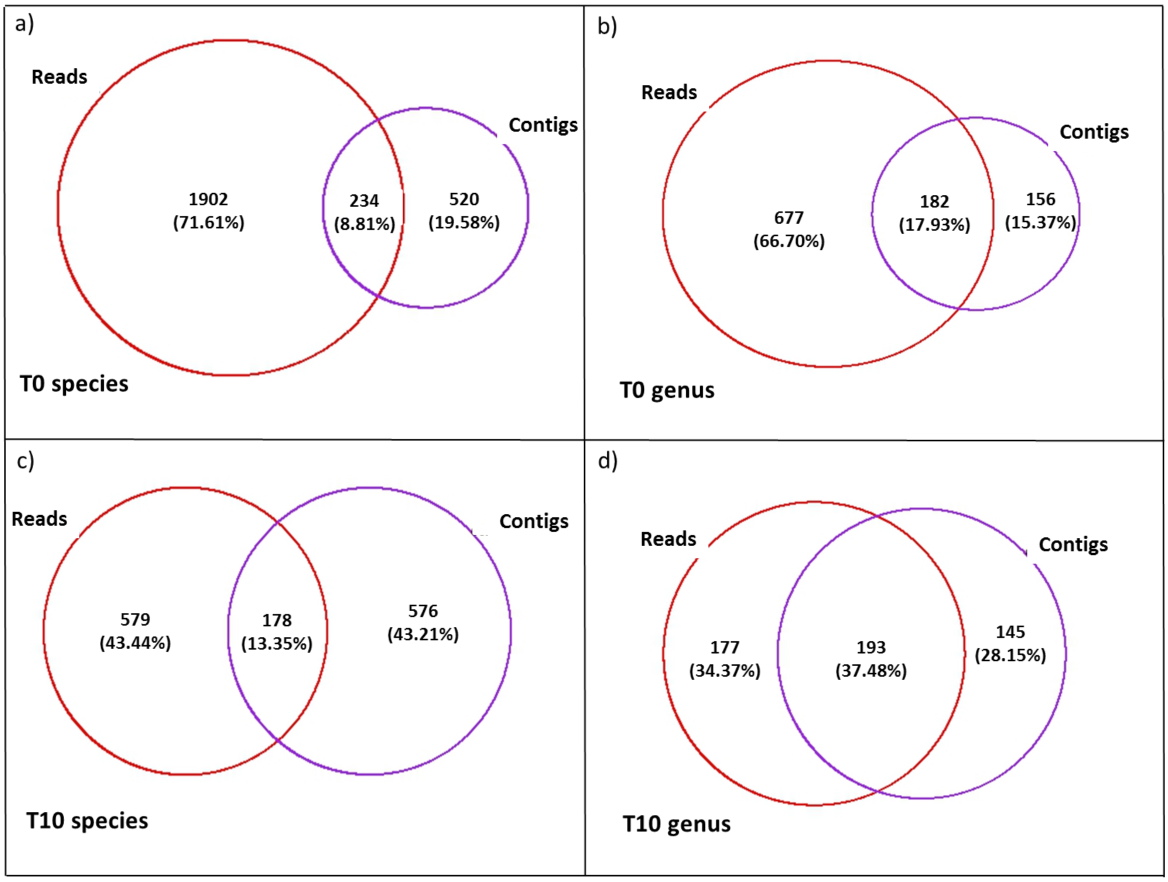


Figure 3.9: Presence absence analysis of <0.45µm fraction: comparison of phylotypes at the level of species (a, c) and genus (b, d) using a subsample of reads (R1) versus contigs at T0 (a, b) and T10 (c, d).

The Refseq annotation (Supplementary Figure 4, Table 3.3) produced an output highly dominated by Actinobacteria (47.30%) and Proteobacteria (38.20%). Specifically, the order Microcroccales made up 40.69% of sequences from the genus *Microbacterium* being the most representative at 33% of all the bacteria. The Proteobacteria could be further subdivided into the classes Alphaproteobacteria (36.51%), Gammaproteobacteria (1.05%) and

Betaproteobacteria (0.56%). The class Alphaproteobacteria was dominated by the order Sphingomonadales (33.25%), with the genus *Erythrobacter* representing 24% of all the contigs, for which one CDS matched a 16S gene (data not shown). Eukaryotes were represented in 1.35% of the metagenomic fraction and were dominated by the family Phaeophyceae (87%) with genus *Ectocarpus* as the major representative (57%), and 19% by the family Laminariaceae, where *Saccharina* was most the common genus. Metazoa constituted only 0.07% of the eukaryotes (Table 3.3, Supplementary Figure 4).

The viral contigs (13.77%, Figure 3.8b) were then annotated using a curated Virus db (Table 3.4 & Supplementary Figure 5). The virome was dominated by sequences mapping to the order *Caudovirales* (59%) comprising the families *Myoviridae* (26%), *Siphoviridae* (22%) and *Podoviridae* (10%) in respective order of abundance (Table 3.4 & Supplementary Figure 5). The NCLDVs (28%) represented the second major group, with the families *Phycodnaviridae* (13%) and *Mimiviridae* (8%) as the main representatives.

Table 3.4: Contigs (<0.45 µm fraction, permeate) annotation using the Virus database. Three top phylotypes per family showed.

Order	Family	Species
Caudovirales (59%)	*Myoviridae* (26%)	*Synechococcus phage S-SM2* (2%) *Phrochlorococcus phage P-SSM2* (2%) *Bacillus phage 0305 phi8-36* (1%) *Synechococcus phage S-SSM7* (1%) *Enterobacteria phage vB_KleM-RaK2* (1%)
	Siphoviridae (22%)	*Enterobacteria phage HIK630* (2%) *Synechococcus phage S-SKS1* (1%) *Microbacterium phage Min1* (1%) *Bacillus phage SPbeta* (1%)
	Podoviridae (10%)	*Planktothrix phage PaV-LD* (6%) *Bordetella phage BIP-1* (0.8%) *Cellulophaga phage phi14:2* (0.4%)
NCLDv (28%)	*Phycodnaviridae* (13%)	*Ectocarpus siliculosus virus 1* (2%) *Paramecium bursaria chlorella virus 1* (2%) *Paramecium bursaria chlorella virus MT325* (1%)
	Mimiviridae (8%)	*Acanthamoeba polyphaga moumouvirus* (3%) *Cafeteria roenbergensis virus BV-PW1* (2%) *Acanthamoeba polyphaga mimivirus* (2%)
Unassigned (12%)		*Megavirus lba* (2%) *Paramecium bursaria Chlorella virus AR158* (1%) *Phaeocystis globosa virus 12T* (1%) *Flavobacterium phage 6H* (1%)

3.3.4 Composition of microbiota in <0.45µm fraction versus >0.45µm fraction

To understand if the prokaryotes and eukaryotes identified in the permeate (<0.45µm) consisted of environmental DNA in the form of debris or vesicles from extant bacteria and eukaryotes present in the water column, stable free DNA, or small bacteria that passed through the filter, presence-absence analyses were run to compare the presence of microbiota in the <0.45µm fraction vs the >0.45µm for each threshold (Figure 3.10). Comparisons were also run at the genus level, or, when the genus annotation was not available for the classification, the higher taxonomic level was utilised. Very little overlap was observed across all levels of stringency (Figure 3.10). The genus *Phaeodactilum* (Supplementary Table 4), shared between all datasets at T0, disappeared when singletons were removed (Figure 3.10b). Commonalities between eukaryotes and prokaryotes showed the presence of chloroplasts and mitochondria in the prokaryotic fraction with genera shared for 1.24% at T0, 0.83% at T1 and 0.45% at T10 (Figure 3.7). When the threshold T1 was applied, it caused the removal of unusual genera such as *Cicer*, *Cucumis*, and *Porphyridium*, whilst genera such as *Chlorella*, *Chroomonas*, *Karlodium* and *Pedinomonas* disappeared with T10 threshold (Supplementary Table 5).

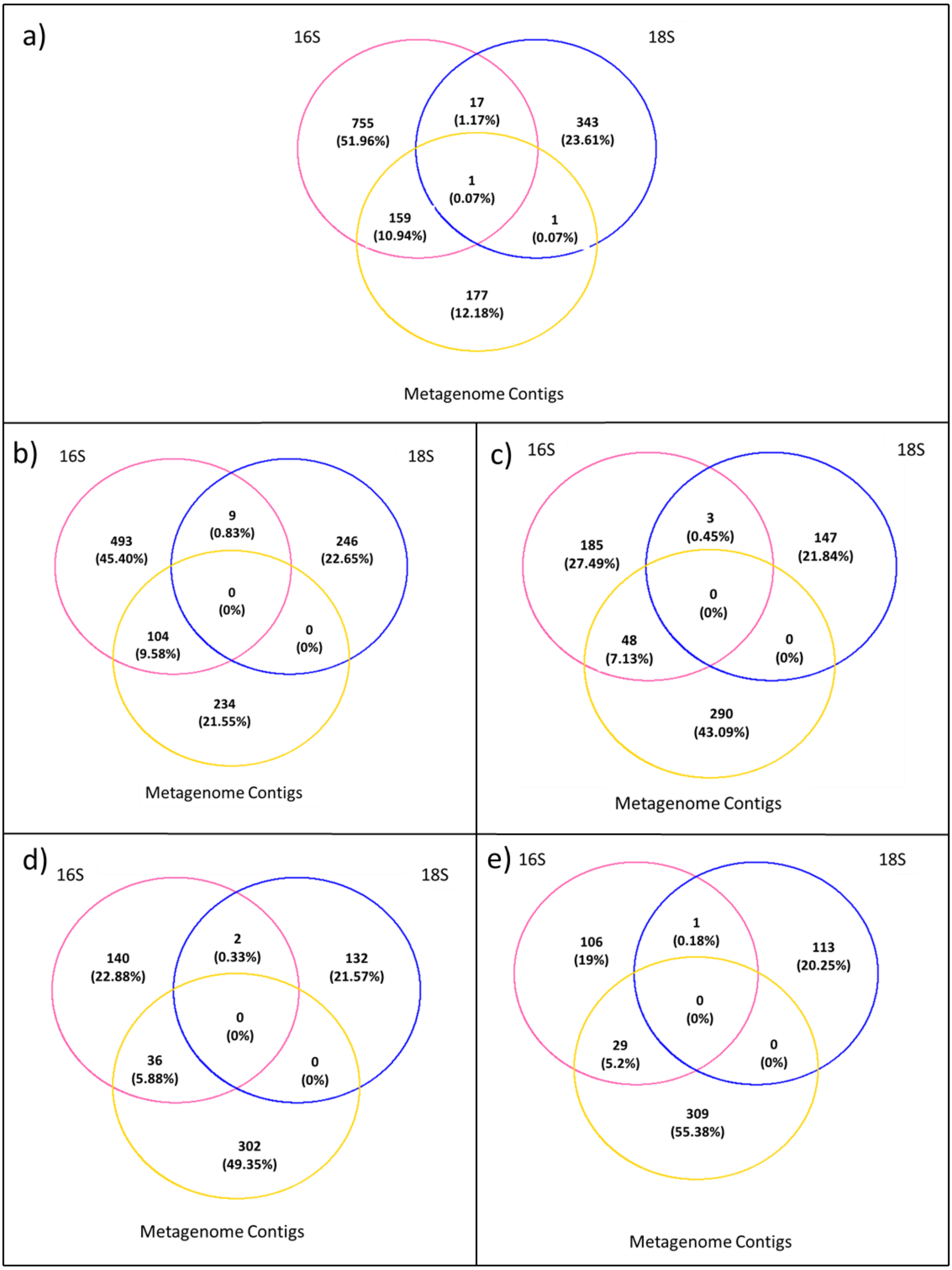


Figure 3.10: Presence-absence analysis between the >0.45µm fraction (prokaryotes and eukaryotes) and the permeate (<0.45µm). a) T0: Metagenomic contigs, prokaryotes, eukaryotes; b) T0: Metagenomic contigs, T1: prokaryotes, eukaryotes; c) T0: Metagenomic contigs, T10: prokaryotes, eukaryotes; d) T0: Metagenomic contigs, T10-R1: prokaryotes, eukaryotes; e) T0: Metagenomic contigs, T10-R2: prokaryotes, eukaryotes.

3.4 Discussion

Describing and studying the hosts, prokaryotes and eukaryote assemblages, alongside their viruses can help improve our understanding on the roles of the microbiome in a holistic way. For this reason, various ocean expeditions were launched to study microbial diversity in its complexity, including different trophic levels and various ecosystem components in a more comprehensive way.

The sampling of microbes in the marine environment has to take into consideration various aspects such as its patchiness (Cao *et al.*, 2002; Deacon, 1982; Frederickson *et al.*, 2003; Seymour *et al.*, 2006) and the fact that this environment changes rapidly, both in time and space. Fingerprint profiles of the marine environment have shown the absence of significant differences in richness when utilising from 10 to 1000ml of water (Dorigo *et al.*, 2006) as well as low variability of the community structure when utilising more than 50ml (Ghiglione *et al.*, 2005). In this study a smaller volume of water (250ml) was used for a sequencing based approach on all three microbial components (prokaryotes, eukaryotes and viruses). In addition, PCR replication provided further confidence when establishing both dominant and rare taxa (chapter 2). The use of four levels of stringency allowed those apparent OTUs produced by sequencing errors and/or contamination to be disregarded. The application of different thresholds sequentially reduced the numbers of phylotypes. The removal of singletons resulted in the reduction of the overall phylotypes by around 700, while retaining over 99% of the reads. This step removed sequences of terrestrial origin (e.g. *Nicotiana* and *Cicer*), which are not expected to occupy the marine microbiome. Notwithstanding that singleton removal is a common practice (Behnke *et al.*, 2011; Zheng *et al.*, 2016; Reeder and Knight, 2009), researchers often retain these taxa under the label of “rare” microbiome members. This study demonstrates that many of these are in fact artifacts, such as sesame and tobacco, rather than true rare species. When singletons are removed in conjunction with replication of PCR reactions, novel to this and the previous study presented in Chapter 2, a more stringent and precise description of the microbiota present in the environment can be obtained. This latter filtering step (T1 on the three replicates combined) allowed the identification of 23 thousand OTUs for the prokaryotic dataset and three thousand for the eukaryotic dataset grouping 834 and 346 as the lowest level of assigned taxa respectively. The further application of a more stringent threshold, i.e. a phylotypes was considered present with at least 10 reads in each PCR replicate, meant that the rare microbiota (i.e. *Chlorella*, *Pedinomonas*, *Marinobacter* and *Oceanicaulis*) were not included in the final dataset. Therefore the removal of singletons, here described as level T1, will be recommended for future studies, allowing to mantain less abundant organisms whilst removing artifacts and sequencing errors from the final dataset.

Bacterial composition at the location analysed by Tara Oceans expedition (station 64), based 548 nautical miles from station S1, showed high abundance of Alphaproteobacteria followed by Cyanobacteria, Gammaproteobacteria and Bacteroidetes (Sunagawa *et al.*, 2015). A similar microbial composition was found in sample from station S1, although it detected the dominance of Cyanobacteria followed by Alphaproteobacteria, Gammaproteobacteria and Bacteroidetes. Eukaryotes collected from Tara Oceans station 64 were dominated by the pico-nanoplankton, the Alveolata (Dinophyceae and Syndiniales clade MALV-I-II, the latest reclassified as Syndiniales groups I and II (Horiguchi, 2015)), followed in abundance by “other protists” (de Vargas *et al.*, 2015); with station S1 was also dominated by Alveolata (Dinophyceae and Syndiniales). It is possible to hypothesise that the variation in composition from station S1 (from this study) and Tara Oceans’ station 64 can be attributed to sampling different water body masses, as well as different sampling seasons; the Tara Oceans survey sampled in the southern hemisphere winter (July 2010), while samples from station S1 were collected in the southern hemisphere summer (February 2012). Given these differences, it is remarkable how similar the microbial communities were, especially when considering the application of vastly different sampling protocols adding confidence to the description of the microbiome in the region and corroborated the microbial paradigm that "everything is everywhere".

Analysis of the metagenomic fraction from the 0.45µm permeate showed that annotation based on contigs led to a more robust description of diversity. The majority of the metagenomic data (86%) did not match any viral genomic region in our curated virus database, this was also reported in previous studies, i.e. 55% (Brum *et al.*, 2013a), 91.4% average (Angly *et al.*, 2006), 88% (Williamson *et al.*, 2012) 64.48% (Breitbart *et al.*, 2002). Marine viral metagenomics or metabarcoding studies currently apply various biomass or volume concentration methods before the extraction of DNA for sequencing (Angly *et al.*, 2006; Hurwitz and Sullivan, 2013; Williamson *et al.*, 2012; Chow *et al.*, 2015; Brum *et al.*, 2015b). Such studies that applied to the area of interest of this study, reported on the dominance of members of the order *Caudovirales*. This dataset similarly report that the latter was the most dominant viral taxa in this environment. Members of the family *Phycodnaviridae* were the major viral group identified for the NCLDVs, followed by members of the family *Mimiviridae* in both this and previous data available for this station (Williamson *et al.*, 2012). Importantly, this study demonstrated that a similar description of virus diversity is achievable from only 250ml of seawater. The high abundance of prochlorococcus and synechococcus phages can be correlated with the presence of cyanobacterial genera such as *Synechococcus* and *Prochlorococcus.* Both co-occurred and dominated the prokaryotic dataset with 30% and 9% of the sequences. NCDLVs, such as members of the families *Phycodnaviridae* and *Mimiviridae*, were surprisingly correlated with the diatoms and dinoflagellates. These taxa, which constituted more than 90% of the eukaryotic dataset, are considered the most widespread microbes on Earth and are known to be routinely infected by RNA viruses (Nagasaki, 2008).

Nevertheless, studies are showing that dinoflagellates are also infected by NCLDVs (Nagasaki, 2008; Nagasaki *et al.*, 2006; Correa *et al.*, 2013), suggesting undescribed host-virus relationships between dinoflagellates and NCLDVs.

Various marine microbial research programs including the Global Ocean Sampling expedition and Tara Oceans expedition have recently surveyed the oceans with the aim of characterising and increasing our knowledge of microbial diversity (Rusch *et al.*, 2007; Williamson *et al.*, 2012; Hurwitz and Sullivan, 2013; Sunagawa *et al.*, 2015; Angly *et al.*, 2006). For these projects, sampling the host and viral fraction simultaneously has been a significant challenge, and the viral fraction has rarely been associated with the host community. Nonetheless, the Tara Oceans campaign provided data for the description of the whole microbiome, outlining the diversity and complexity of bacteria, eukaryotes and viral taxa (Brum *et al.*, 2015b).

Here, an alternative method was utilised to allow the collection and identification of not only the viral fraction, but also the prokaryotic and eukaryotic communities associated with the same water mass. The type of sampling conducted in this study, which excludes water concentration protocols, allowed the characterisation of the viruses present in the <0.45µm permeate and also to look for any associated cell-derived exudates (also referred to as eDNA, or free DNA) in the water collected. The comparative analyses of the two sampled size fractions revealed that bacteria and eukaryotes identified in the environment were not the source of all the eDNA in the sample, since on average 10% were in common with T0 and T1 thresholds. The likely explanation for the source of this eDNA could be either the presence of viruses carrying host genes, since host genes have been identified in virus isolates (Millard *et al.*, 2009), or the presence of small bacteria (>0.45µm). The latter included genera identified in both datasets such as *Pseudomonas*, *Flavobacterium*, *Serratia* and *Vibrio,* which are known to pass through 0.45µm filters (Tabor *et al.*, 1981; Hasegawa *et al.*, 2003). Nine coding sequences of the <0.45µm fraction had BLAST hits with 16S rRNA genes six of which corresponded to *Microbacterium* (data not shown), and represented the main genera identified in this fraction. Furthermore it has been shown that, in adverse conditions, *Microbacterium* undergo a size reduction, which would allow it to pass through 0.45µm filters (Chicote *et al.*, 2005; Iizuka *et al.*, 1998). However, since viruses can acquire host genes through horizontal gene transfer, and a large proportion of genetic material with unknown identity was also described; it is thus feasible to hypothesise that viruses and not bacteria are the likely source of this genetic material (Chow and Suttle, 2015; Millard *et al.*, 2009).

To conclude, this study proposes an alternative method to evaluate the microbiome of any aquatic environment. The marine microbial world, which was previously overlooked, can now be fully explored thanks to recent advancements in next generation sequencing. Taking advantage of these, this study exploits the use of smaller water volumes to characterise microbial diversity, showing that 250ml of water can represent the current description of microbial diversity. For the first time, the use of replication and different filter/threshold were applied to better discriminate genuine and rare phylotypes over sequencing noise. Finally, this study opens the door to a more integrated approach of oceanographic sampling, thereby allowing for better parameterisation of global biological models.

Chapter 4: From protists to viruses, is everything everywhere? The Antarctic Polar Front and microbial distribution in the southern hemisphere oceans

4.1 Introduction

In the past 15 years the world's oceans have been explored far and wide (Figure 1.4) to improve the understanding of marine microbial communities. Various expeditions such as the Global Ocean Sampling (Rusch *et al.*, 2007), the Tara Ocean Expedition (Sunagawa *et al.*, 2015) and the Malaspina (Laursen, 2011) are contributing to new discoveries on microbial diverstiy as well as the presence of microbial spatial patterns (reviewed in Green and Bohannan, 2006) and the structuring of microbial diversity due to both geography and environment (Williamson *et al.*, 2008; de Vargas *et al.*, 2015; Malviya *et al.*, 2015). However these remarkable global efforts had, and continue to have, difficulties. The marine environment is extremely variable with fluxes and currents that generate inconsistency in time and space (Cao *et al.*, 2002; Deacon, 1982; Frederickson *et al.*, 2003; Seymour *et al.*, 2006) with consequent difficulties in the collection of the same sample in different intervals as instead happen, for example, in the collection of soil. Therefore this renders the determination of standard sampling very difficult (Zinger *et al.*, 2012). Nonetheless, perceived low variability in the community structure has meant that small volumes of water (50ml) are considered sufficient for diversity assessments (Ghiglione *et al.*, 2005). This was validated by another study where no significant differences in richness were observed when comparing DNA fingerprinting profiles from 10ml to up to 1L of water (Dorigo *et al.*, 2006). The use of smaller water volumes was also possible due to the improvement in sequencing chemistries and technologies (Hoeijmakers *et al.*, 2011; Marine *et al.*, 2011). Indeed, in Chapter 2 and Chapter 3 it was shown that a small volume of water, 250ml, can reveal coherent bacterial community structures, comparable to studies with larger volumes sampled. Furthermore it was shown that the most abundant phylotypes present in a given sample can be fully described. The use of a small volume of water collected through Conductivity, Temperature and Density (CTD) rosette sampler allowed the sampling of prokaryotes, eukaryotes, viruses and environmental DNA (eDNA) at the same time from a common body of water (as shown in Chapter 3). This is of great importance when applyied to the study of microbial diversity in relation to the understanding of their longitudianal distribution. In fact, in the marine environment, microbial distribution is not uniform and shows variation in both vertical and horizontal dispersal patterns (Salcher *et al.*, 2011). Therefore the presence of fronts such as the Antarctic Polar Front (APF), characterised by intense currents and a strong thermocline (Eastman, 1993; Thornhill *et al.*, 2008), could create a barrier to the microbial genetic flow as shown for some eukaryotes (Thornhill *et al.*, 2008; Shaw *et al.*, 2004; Hunter and Halanych, 2008). This specific front separates, two very difference oceanic systems, the Indian Ocean and the Souther Ocean; the first is characterised by upper warm and salty water (Donners and Drijfhout, 2004; Beal *et al.*, 2011), the second by colder water with evidence of iron limitation (Popova *et al.*, 2000).

In this Chapter the full microbial diversity of virus, protists and bacteria from samples collected across two oceanic systems separated by the Antarctic Polar Front (APF) will be described. Amplicon sequencing was used to characterise the microbiota present; specifically the V4 region along the 16S rRNA gene and the V9 region of the 18S rRNA gene were used to analyse the prokaryotic and eukaryotic community, respectively. Furthermore the viral fraction, together with the environmental DNA (eDNA), was analysed using the metagenome shotgun Illumina sequencing approach. Specifically, eDNA represents DNA that have been released into the environmental (i.e. water, soil etc.) without isolating it directly from a target organism. It therefore is composed of a mixture of DNA derived from cellular debris or released DNA from biota living in that environment (Taberlet *et al.*, 2012). In the past, eDNA has been used as a tool to determine whether an invasion has taken place (Dejean *et al.*, 2012) or to track an endangered species (Ikeda *et al.*, 2016). Therefore it has been previously proposed that eDNA could be used as a monitoring tool (Valentini *et al.*, 2016); here I will determine whether the biota can be monitored by using the eDNA/virus fraction.

4.2 Materials and Methods:

4.2.1 Sampling

Six samples were collected during the Great Southern Coccolithophore Belt expedition (GSCB-cruise RR1202: http://www.bco-dmo.org/project/473206) (Figure 4.1, Table 4.1). Stations S1 and S2 were located in the South-West Indian Ocean, stations S3 and S4 in the Southern Ocean, and stations S5 and S6 in the South-East Indian Ocean (Figure 4.1). The locations of the sampling stations along the transect from the south Indian Ocean to the Southern Ocean were mapped using RgoogleMaps_1.2.0.7 (Loecher and Ropkins, 2015) under R version 3.3.0 (2016-05-03) (Figure 4.1, Table 4.1). Samples were collected and the filters were used for the DNA extraction as described for station S1 in chapter three.

Table 4.1: Sampling location information. Coordinates are provided in decimal degrees. Sampling depth refers to the depth of the chlorophyll-*a* maximum. Lat: latitude. Lon: longitude. †: adjusted µmol/Kg SW. ‡: µmol/L. * adjusted µatm

Station	Area	Date of collection (dd/mm/yy)	Lat	Lon	Sampling depth m	T (ITS) ºC	pH	Salinity (PSS-78)	CO3 †	CO2 †	NO_3 ‡	PO_4 ‡	NO_2 ‡	NH_4 ‡	pCO_2 *	HCO_3 †
S1	SW Indian Ocean	20/02/12	-38.315	40.958	5	20.83	8.08	35.567	211.99	11.15	0.59	0.21	2.9	0	354.35	1815.32
S2	SW Indian Ocean	22/02/12	-35.507	37.458	49.089	19.98	8.08	35.483	207.07	11.34	NA	NA	NA	NA	351.52	1824.18
S3	Southern Ocean	06/03/12	-57.598	76.508	41.855	1.38	8.05	33.913	97.67	22.21	27.44	1.87	42.8	0.24	371.72	2027.16
S4	Southern Ocean	06/03/12	-58.710	76.890	40.93	1.24	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
S5	SE Indian Ocean	17/03/12	-39.475	108.935	44.978	16.23	8.11	35.081	189.62	11.87	1.93	0.28	0.6	0.02	329.94	1844.2
S6	SE Indian Ocean	19/03/12	-42.082	113.400	60.55	12.95	NA	34.822	NA	NA	6.26	0.63	0.6	0.4	NA	NA

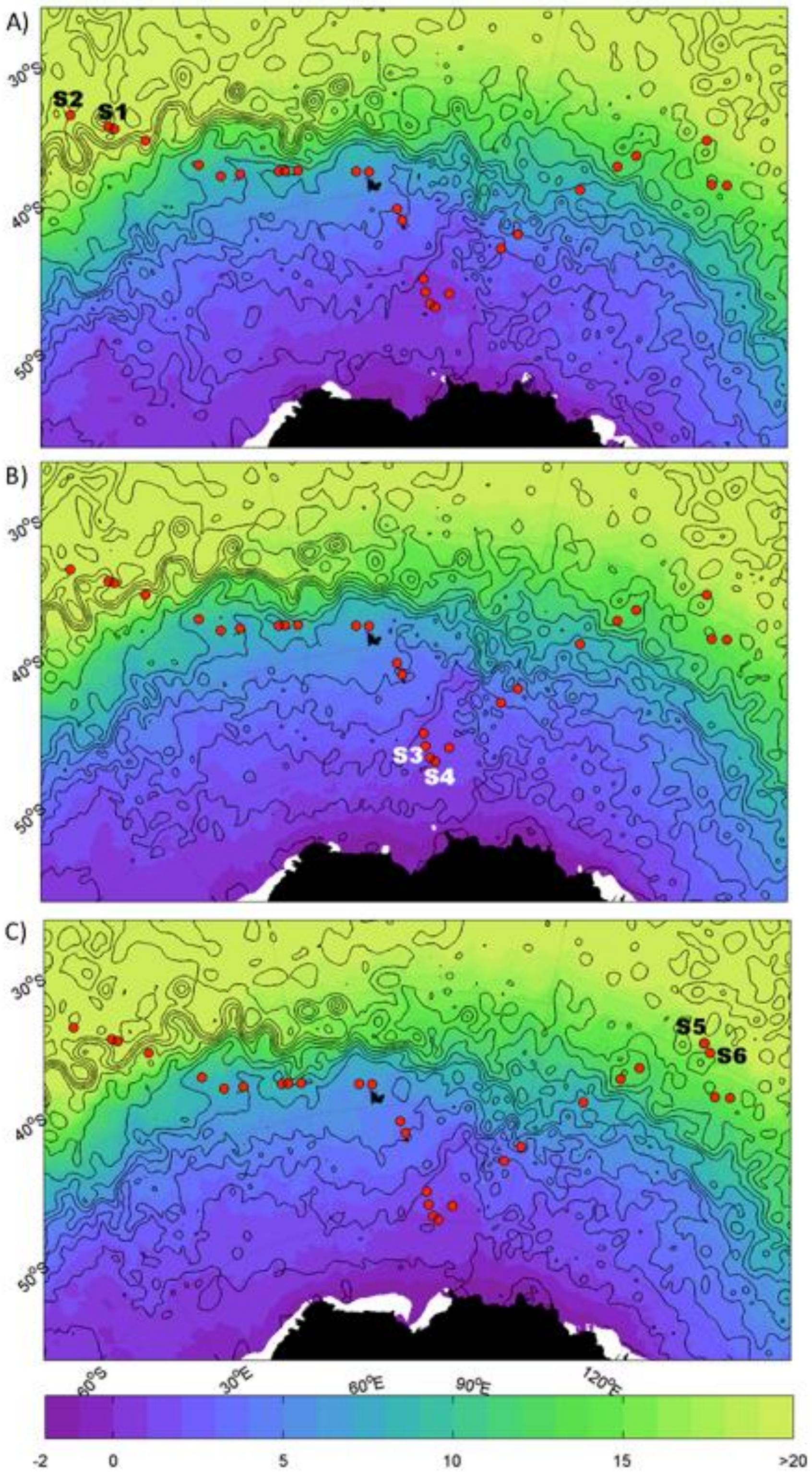


Figure 4.1: Map of sampling sites. Sample locations for the Southern Hemisphere, combined with a background showing current flows and sea surface temperature (SST in °C). The black lines are dynamic height and the closed contours show eddies. The main current flow, the Antarctic Circumpolar Current (ACC) is reported by black arrows. A) refers to 15/02/2012; B) refers to 29/02/2012 and C) refers to 14/03/2012.

4.2.2 Preparation and sequencing of the >0.45µm fraction and <0.45µm fraction (virome)

For the prokaryote community composition analysis, the V4 region of 16S ribosomal RNA gene was amplified using the universal primer pair 515F/806R and Illumina tagged primers (Caporaso *et al.*, 2012a) (Appendix I). Eukaryotes were characterized using the 18S ribosomal RNA gene, using primer pair 1391F/EukB and Illumina tagging to amplify the V9 region (Stoeck *et al.*, 2010) (Appendix I). Real time PCRs, run for each sample to determine the mid-exponential threshold of each reaction, amplicon preparation and sequencing followed the same protocol as described in Chapter 3. The whole 50ml filtrate, hereafter described as permeate, was subjected to a nucleic acid extraction procedure, one sample per station, and sequencing followed the same protocol as described for sample S1 in chapter three.

4.2.3 Bioinformatics pipeline

Bioinformatics pre- and post-processing followed the pipeline described in Chapter 3 (Figure 3.2) using both the Bio-Linux 8 system at the Marine Biological Association of the UK as well as the facilities provided by the University of Cape Town's ICTS High Performance Computing team: http://hpc.uct.ac.za. Sequencing information for both amplicon dataset (prokaryotes V4-16S amplicons and eukaryotes V9-18S amplicons, Table 4.2) and the permeate (<0.45µm permeate, metagenome Table 4.3) show both raw and cleaned reads as well as contigs retrieval. Community composition was visualised using Krona tools (Ondov *et al.*, 2011) after taxonomic assignments. Venn diagrams were created using the R package VennDiagram_1.6.17 on R version 3.3.0 (2016-05-03) to determine the number of shared OTUs and phylotypes among the sequencing methods used.

Table 4.2: Stepwise processing of prokaryote (16S rRNA) and eukaryote (18S rRNA) sequences. The raw sequence reads (a) were first preprocessed to remove adapters (b) and then trimmed and filtered (c), before OTUs were assigned (d). Singletons were removed (e) and the final OTUs per sample assigned (f). The number of unique OTUs was calculated for 16S and 18S-derived datasets. *16S removal of chloroplasts and mitochondria. ** Unique OTUs that are recovered in the dataset across all six stations.

Dataset	Sample	(a) Raw Reads	(b) First Preprocessing	(c) Cleaned reads (125 b)	(d) OTU	(e) Reads after Singletons Removed* (T1)	(f) OTU – Singletons removed* (T1)
16S	S1a Rep1	1,331,542	773,343	741,033	20,381	705,707	10,281
	S1a Rep2	1,695,911	1,161,634	1,117,576	30,642	1,072,726	15,398
	S1a Rep3	1,626,930	863,867	841,639	24,756	813,380	12,626
	S2 Rep1	983,760	443,622	437,790	17,141	374,459	6,775
	S2 Rep2	1,458,024	627,698	619,255	25,586	525,969	10,078
	S2 Rep3	1,550,314	646,303	637,701	25,208	543,433	9,683
	S3 Rep1	1,491,664	622,030	609,658	18,305	265,157	6,027
	S3 Rep2	1,409,872	795,524	781,413	19,487	353,430	6,170
	S3 Rep3	1,754,942	878,836	864,910	24,344	502,820	9,382
	S4 Rep1	974,224	438,389	434,686	15,668	276,871	5,349
	S4 Rep2	1,609,312	721,401	714,793	20,437	422,624	7,078
	S4 Rep3	1,468,624	795,217	788,622	24,338	567,976	9,182
	S5 Rep1	1,497,998	805,139	785,754	27,469	557,937	10,561
	S5 Rep2	838,777	725,672	706,520	27,206	510,315	10,192
	S5 Rep3	1,253,530	725,301	708,433	28,896	518,912	11,215
	S6 Rep1	1,477,596	664,590	659,890	16,141	205,402	5,563
	S6 Rep2	1,695,898	761,187	755,509	18,741	239,579	6,271
	S6 Rep3	771,891	696,673	691,459	17,978	240,741	6,058
Total 16S				**12,896,641**	**133,550****	**8,697,438**	**48,923****
18S	S1a Rep1	1,529,536	305,949	223,814	2,972	222,556	1,714
	S1a Rep2	1,614,464	374,041	275,201	3,271	273,710	1,780
	S1a Rep3	1,695,911	419,375	308,208	3,470	306,574	1,836
	S2 Rep1	1,258,768	269,903	179,753	4,499	177,824	2,735

S2 Rep2	1,626,930	528,707	354,840	5,454	352,118	3,506
S2 Rep3	1,491,664	425,343	286,080	4,401	283,533	3,260
S3 Rep1	1,331,542	417,425	33,738	1,043	33,347	288
S3 Rep2	1,505,002	80,678	7,279	4,369	6,977	163
S3 Rep3	1,685,214	583,156	71,509	4,595	70,284	1,469
S4 Rep1	1,392,915	323,407	38,196	4,664	37,698	382
S4 Rep2	1,393,132	387,326	50,317	6,228	49,754	520
S4 Rep3	1,403,962	389,608	47,930	5,807	47,464	365
S5 Rep1	1,188,018	336,163	202,101	880	199,824	2,222
S5 Rep2	1,799,244	461,542	278,117	1,083	275,184	2,521
S5 Rep3	1,238,172	310,004	186,059	831	183,800	2,142
S6 Rep1	838,777	14,341	9,291	679	8,807	559
S6 Rep2	1,253,530	306,231	195,247	465	193,024	2,146
S6 Rep3	1,284,848	345,184	223,715	2,694	221,307	2,194
Total 18S			**2,971,395**	**30,169****	**2,943,785**	**9,806****

Table 4.3: Stepwise processing of the permeate metagenome. From raw reads to number of contigs assembled using CLC genomic workbench. N50 is calculated by CLC genomic workbench and represents a weighted median statistics on the average assembly which summarise the length of the longest contigs until 50% of the total contigs are reached (https://www.qiagenbioinformatics.com/)**.**

Dataset	Sample	Raw Reads	Reads used for contigs	Number of contigs	Average contigs Length	Smallest contig	Largest contig	N50
Metagenome	S1a	90,672,808	10,036,627	4,962	1,045	240	74442	7239
	S2	16,569,598	16,569,598	35,358	1,060	206	282176	6999
	S3	21,466,152	21,466,152	20,597	1,492	206	388233	3668
	S4	21,840,372	21,840,372	15,844	1,492	230	563674	8321
	S5	14,268,562	14,268,562	18,540	1,312	217	478618	5267
	S6	41,108,086	41,108,086	7,539	2,092	249	1026488	161188

4.2.4 Statistical analysis of the >0.45µm fraction (prokaryotes and eukaryotes)

As described previously Chapter 2 and Chapter 3, the level T1 was chosen to run the analysis on the remaining stations, implying the removal of all singletons (i.e. only one read was present for a defined OTU) in each replicate before running the analysis. Chloroplast and mitochondrial sequences were removed from the prokaryotic dataset prior to the analysis, because they are representative of possible members of the eukaryotic fraction. An R Script (Appendix II) was used to run a number of statistical analyses for 16S and 18S datasets combining functionality of the following R packages: reshape 2_1.4.1, reshape_0.8.5, gclus_1.3.1, GGally_1.0.1, scales_0.4.0, car_2.1-2, picante_1.6-2, nlme_3.1-125, ape_3.4, plyr_1.8.2, amap_0.8-14, gridExtra_0.9.1, ggplot2_2.1.0, clusterSim_0.44-2, MASS_7.3-45, cluster_2.0.3, vegan_2.2-1, lattice_0.20-31, permute_0.8-3, sfsmisc_1.1-0.

Before running any statistical analysis for both the prokaryotic and eukaryotic datasets, the number of reads were normalised to the minimum number present in each dataset, in order to avoid bias caused by different sequencing depths. Alpha diversity was estimated based on species richness derived from OTU richness. Beta diversity was estimated using the *vegan* package, based on the Bray Curtis distance and plotted as hierarchical clustering. Using the full (not normalised) dataset, relative abundance for each group was calculated and plotted using the *ggplot2* package. In order to test if community composition was significantly different between sampling stations Permanova analyses were performed using *Adonis* from the *vegan* package, taking into consideration both temperature and location. To further analyse the community composition Anova was used to determine if the alpha diversity was statistically different between stations; this analyses was performed using the R package *car* using the same two parameters utilised in the Permanova. Finally, the Tukey's post hoc test based on species observed was performed to test if the number of species varied between locations.

4.2.5 Statistical analysis of the 0.45µm permeate (virome)

Due to the lack of replication of the viral sample, Log likelihood ratio statistic was used to test the goodness of fit for two models. The first model (H0) implied that pairwise sampling stations grouped by locations (South West Indian Ocean and Southern Ocean; South West Indian Ocean and South East Indian Ocean; Southern Ocean and South East Indian Ocean) had the same underlying viral distribution. The second model, the alternative hypothesis (H1) assumes that the distribution of viruses depended on the locations. The p-value based on the likelihood ratio was then computed (Appendix II).

Comparison of prokaryotic and eukaryotic amplicons with the metagenome was run through presence absence analyses plotted as Venn diagrams using R package VennDiagram_1.6.17. For the metagenomic fraction the Refseq annotation was used while prokaryote and eukaryote taxonomy was assigned using SILVA. In order to avoid conflicts on species annotation or variation in species names in the different databases, comparative analyses were run at the genus level, or the first available taxonomic level above.

4.3 Results

4.3.1 Prokaryotic diversity and composition in the 0.45 µm fraction

A total of 12.9 million prokaryotic sequences, obtained for all of the six samples, could be clustered into 133,550 OTUs. When singletons, chloroplast and mitochondria were removed a total of 8.7 million sequences could be clustered into 48,923 OTUs (Table 4.2). Of this ~50k OTUs 44.37% were shared across the six locations (Table 4.4) and 1.65% shared across all six stations (807, Figure 4.2a). Specifically 31.17% of the OTUs were unique to the south-west

Indian Ocean, 23.30% present exclusively in the Southern Ocean and 15.04% belonging to the south-east Indian Ocean (Table 4.4, Figure 4.2).

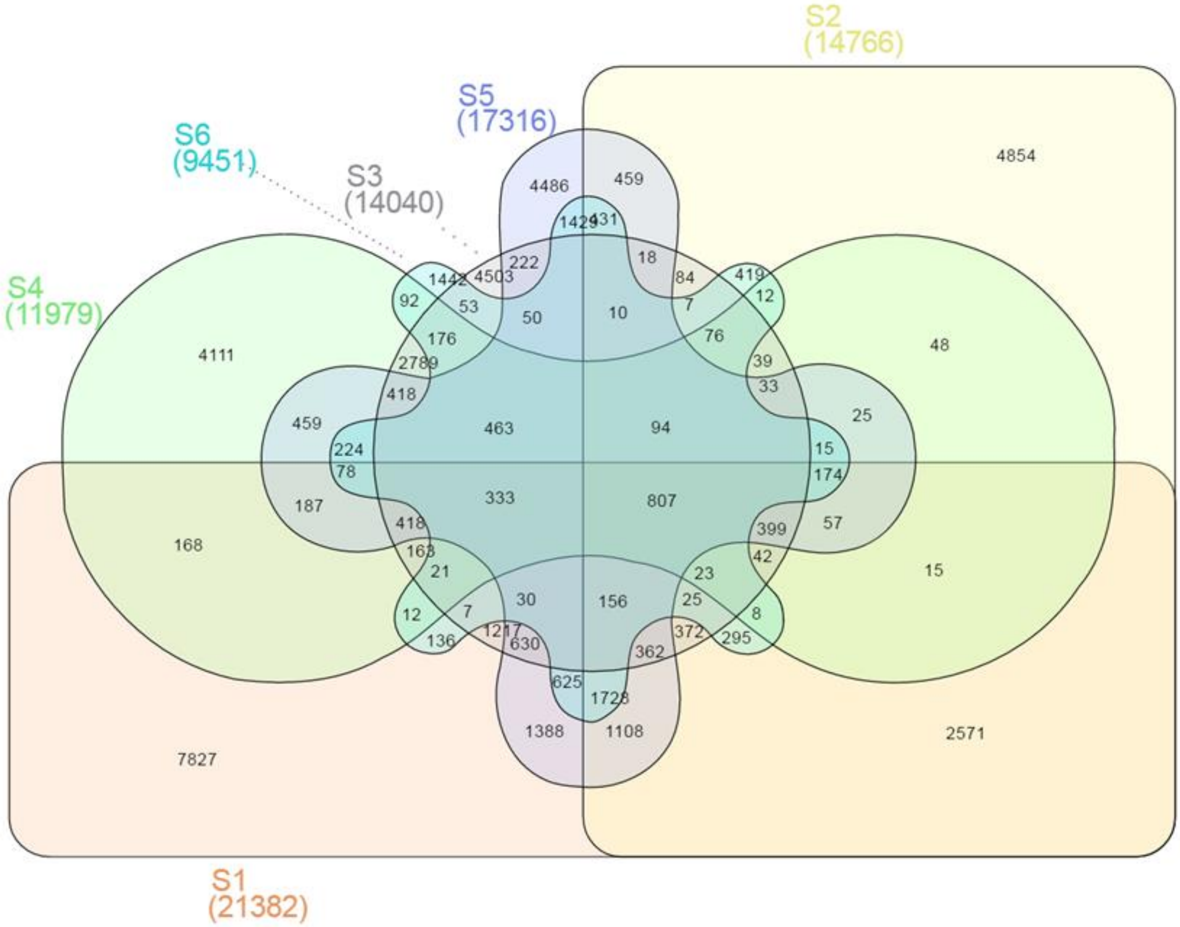


Figure 4.2: Six way Venn diagram for the prokaryotic community based on OTUs per station. Each colour represents a different station. South-west Indian Ocean: station S1 orange and station S2 yellow. Southern Ocean: station S3 gray, station S4 green. South-east Indian Ocean: station S5 blue and station S6 cyan.

The prokaryotic fraction was dominated by bacterial sequences (average 88.34 ±0.08%, min = 79.84% in S1 and max = 97.45% in S6) whereas reads with no annotation represented on average 11.01 ±0.09% of the full dataset (min = 0.66% in S6 and max = 22.59% in S3; Suppl. Tab. 4.1). Archaea were identified in 0.65 ±0.82% of the sequences (min = 0.01% in S4 and max = 1.89% in S6; Suppl. Tab. 4.1).

Table 4.4: Number of unique OTUs found at each station and location.

	16S rRNA				18S rRNA			
	Number of OTU	%	LOCATION	%	Number of OTU	%	LOCATION	%
Unique OTUs S1	7827	16	15252	31.17	1394	14.22	4308	43.93
Unique OTUs S2	4854	9.92			2725	27.79		
Unique OTUs S3	4503	9.2	11403	23.31	98	1	388	3.96
Unique OTUs S4	4111	8.4			221	2.25		
Unique OTUs S5	4486	9.17	7357	15.04	1293	13.19	2578	26.29
Unique OTUs S6	1442	2.95			892	9.1		
Shared OTUs	21709	44.37			3183	32.46		
Total Number of OTUs	48923				9806			

The bacterial fraction was dominated by the phylum Proteobacteria, representing on average 49.55 ±16.59% of the sequences (min = 24.55% in S2 and max = 64.91% in S4; Supplementary Table 6). This group could be further separated into the Gammaproteobacteria (average = 22.93 ±11.64%, min = 8.04% in S2 and max = 32.32% in S3), Alphaproteobacteria (average = 22.88 ±8.2%, min = 11.28% in S2 and max = 36.36% in S4) and Deltaproteobacteria (average = 1.39 ±0.96%, min = 0.24% in S4 and max = 4.11% in S2) (Supplementary Table 6). Cyanobacteria represented the second main phylum, constituting on average 21.34 ±%23.54 with a minimum of 0.03% in S4 and reaching a maximum of 58.86% in S2 (Supplementary Table 6). The third most represented phylum was Bacteroidetes with an average of 12.76 ±8.38% (min = 3.26% in S1 and max = 24.37% in S4), mainly due to a high presence of sequences identified as class Flavobacteriia (average = 12.39 ±8.3%, min = 3.21% in S1 and max = 23.86% in S4).

The three most common bacterial groups varied between stations (Figure 4.3). The south-west Indian Ocean station S1 was characterised mainly by Cyanobacteria, followed by Gammaproteobacteria and Alphaproteobacteria (Figure 4.3, Supplementaty Figure 6). S2 was dominated by Cyanobacteria followed by Alphaproteobacteria and Gammaproteobacteria (Figure 4.3, Supplementaty Figure 7). The Southern Ocean station S3 composition was characterised mainly by Gammaproteobacteria followed by Alphaproteobacteria and Flavobacteriia whilst station S4 by Alphaproteobacteria, Gammaproteobacteria and Flavobacteria (Figure 4.3, Supplementaty Figures 8 & 9). Both stations located in the south-east Indian Ocean, stations S5 and S6, were characterised by a presence of Gammaproteobacteria, Alphaproteobacteria and Cyanobacteria respectively in order of abundance (Figure 4.3, Supplementaty Figures 10 & 11).

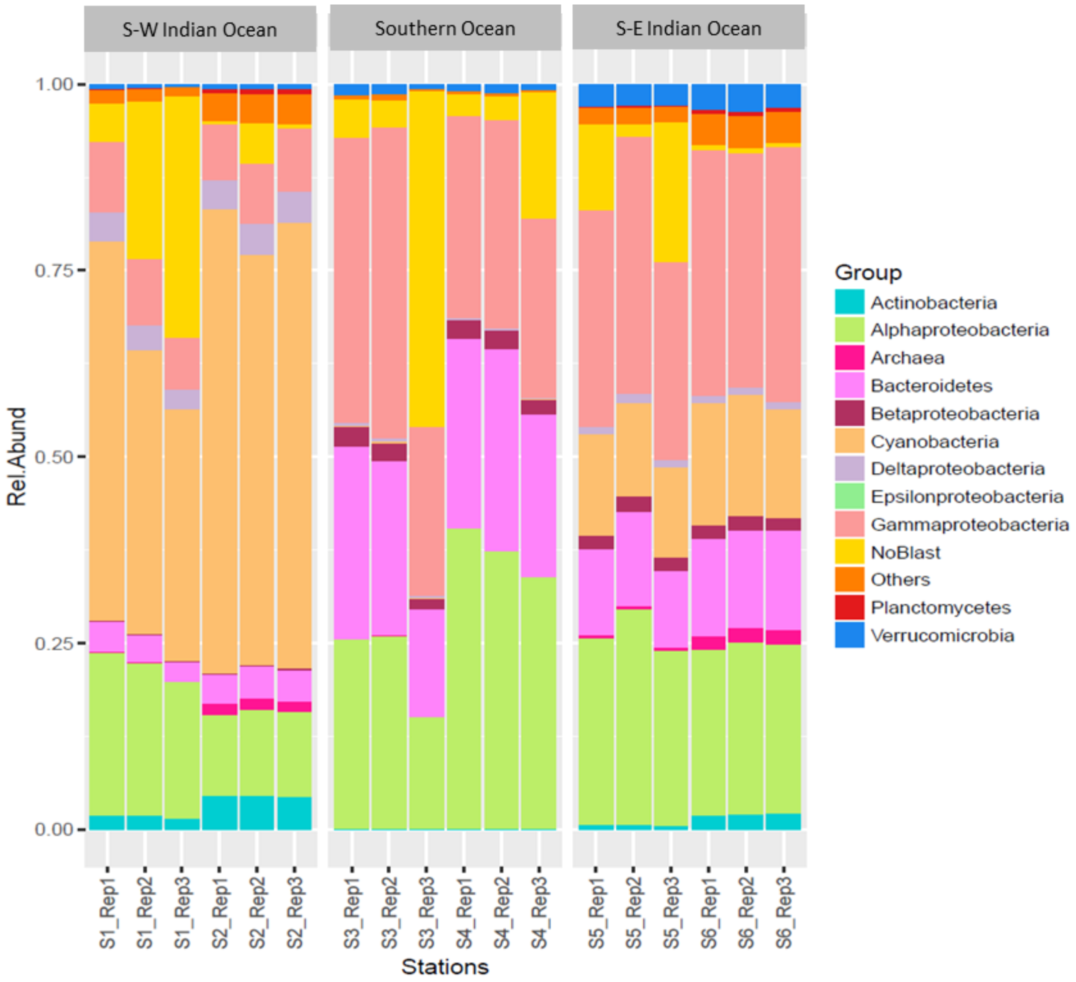


Figure 4.3: Prokaryotic relative abundance composition after singletons, chloroplast and mitochondrial sequences were removed.

4.3.2 Geographic comparison of prokaryotic communities

Bacterial OTU composition differed significantly between the three main locations (Figure 4.3; PERMANOVA $F_{2,12}$ = 64.549, p = 0.001*). Species richness was also significantly different between locations (ANOVA $F_{2,12}$=5.28, p= 0.0227*). A post hoc Tukey's test showed that only the south-west Indian Ocean and the Southern Ocean were significantly different in the number of species to the Southern Ocean (p adj >0.01). Neither the two southern Indian Ocean station (p adj = 0.348) or the south-east Indian Ocean and the Southern Ocean (p adj = 0.213) were significantly different in the number of species. The Bray-Curtis dissimilarity matrix analysed through hierarchical clustering shows that the 6 stations clustered according to the three locations, i.e. stations within the south-west Indian Ocean, south-east Indian Ocean, or Southern Ocean clustered closely together while the distance between these clusters was significant (Figure 4.4).

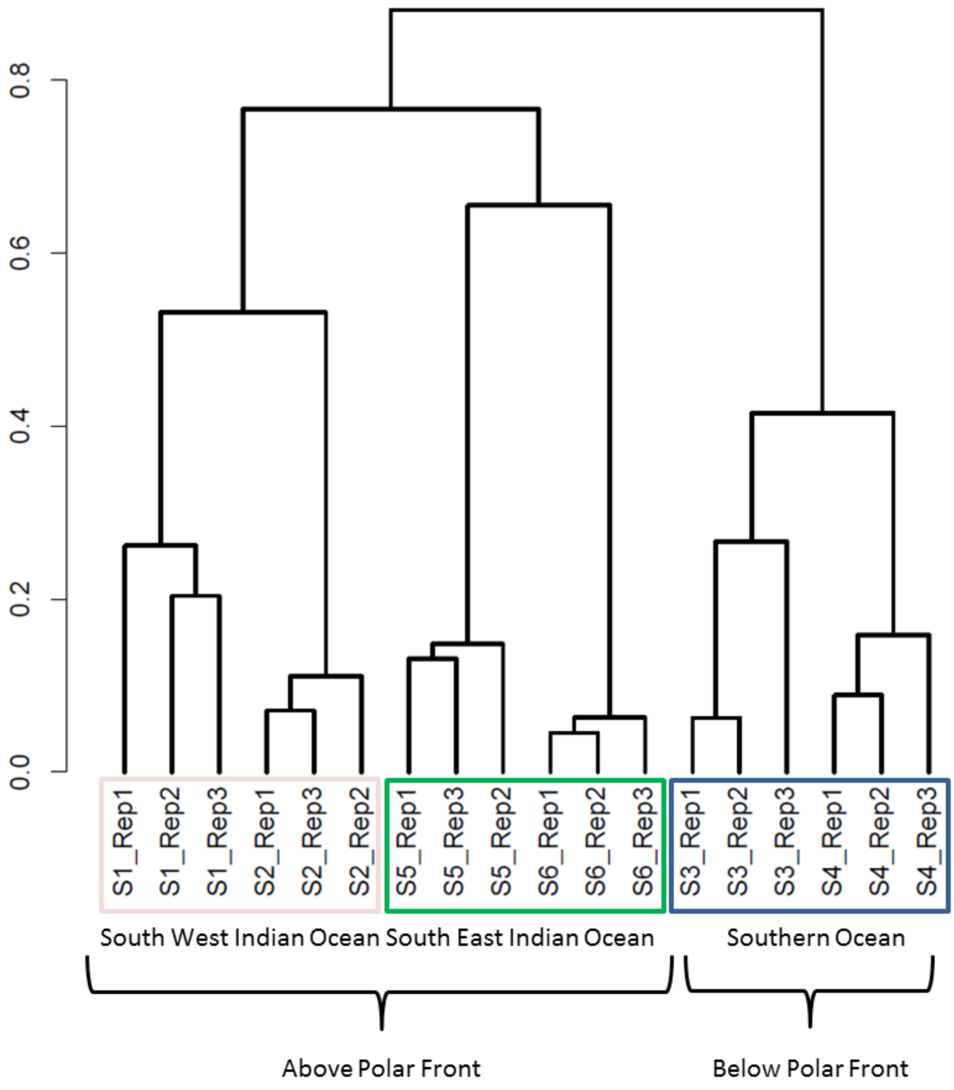


Figure 4.4: Hierarchical clustering based on Bray-Curtis dissimilarity matrix; prokaryotic dataset.

4.3.3 Eukaryote biodiversity and community composition in the 0.45 µm fraction

For the eukaryotic fraction 5.94 million sequences clustered into 30,169 OTUs. After the removal of singletons a total of 5.88 million sequences clustered into 9,806 OTUs (Table 4.2). Of the almost 10 thousand OTUs 32.46% were shared across the six locations and 1.96% shared across the six stations (192, Figure 4.5). Specifically, 43.93% of the eukaryotic OTUs were unique to the south-west Indian Ocean, 3.96% were present exclusively in the Southern Ocean, and 26.29% belonged to the south-east Indian Ocean (Table4.4).

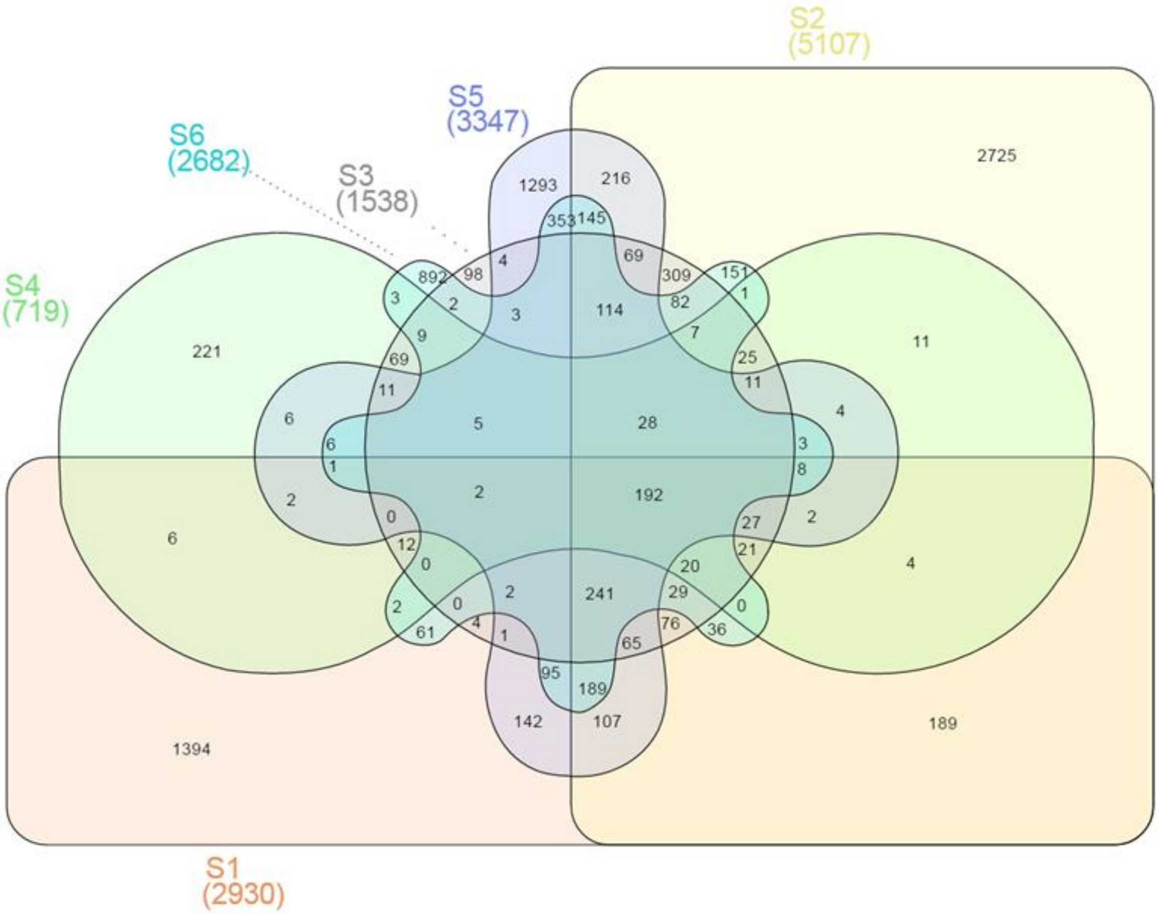


Figure 4.5: Six way Venn diagram for the eukaryotic community based on OTUs per station. Each colour represents a different station. South-west Indian Ocean: station S1 orange and station S2 yellow. Southern Ocean: station S3 gray, station S4 green. South-east Indian Ocean: station S5 blue and station S6 cyan.

The eukaryotic dataset was fully dominated by known sequences, with an average of uncharacterised sequences of only 0.32 ±0.31% (Supplementary Table 7). Eukaryotes were dominated at all stations by the supergroup SAR (Stramenopiles, Alveolata, Rhizaria) representing 85.52 ±9.80% of all sequences (min = 69.99% in S4 and max = 93.27% in S6; Figure 4.3a). For this supergroup the major representative was the superphylum Alveolata (average = 83.52 ±9.80%, min = 68.98% in S4 and max = 91.66% in S1) followed by Rhizaria (average = 1.72 ±0.91%, min = 0.41% in S4 and max = 2.84% in S2). The second main group was represented by the division Haptophyta (average = 9.42 ±9.80%, min = 1.15% in S2 and max = 27.17% in S4) with Prymnesiophyceae as the main representative (average = 9.10 ±11.10%, min = 1.15% in S2 and max = 27.17 in S4) with *Phaeocystis globosa* representing 26.4% of S4 and 17.7% of S3 while less than 1% at the remaining stations (Supplementary Table 7).

The three most abundant eukaryotic groups (Figure 4.6), annotated per station with Silva level four of taxonomy (L4), were as follows. South-west Indian Ocean S1 was characterised by Protoalveolata (43.86%), Dinoflagellata (41.35%) and Ciliophora (3.40%) all belonging to the group SAR/Alveolata (Figure 4.6, Supplementary Figure 2). S2 was represented by Dinoflagellata (42.14%), Protoalveolata (40.73%) and Ciliophora (3.12%; Figure 4.6, Supplementary Figure 12). Southern Ocean station S3 was dominated by sequences from Dinoflagellata (37.57%), Protoalveolata (33.80%) and Haptophyta-Prymnesiophiceae-*Phaeocystis* (18%; Figure 4.6, Supplementary Figure 13), while the second polar station (S4) was characterised by Dinoflagellata (36.47%), Haptophyta-Prymnesiophiceae-*Phaeocystis* (26%) and Protoalveolata (22.57%; Figure 4.6, Supplementary Figure 14). Station S5 in the south-east Indian Ocean was composed of Dinoflagellata (43.78%), Protoalveolata (35.68%) and Ciliophora (4.71%; Figure 4.6, Supplementary Figure 15) whereas S6 was dominated by

Protoalveolata (57.73%) followed by Dinoflagellata (27.58%) and Ciliophora (3.21%; Figure 4.6, Supplementary Figure 16).

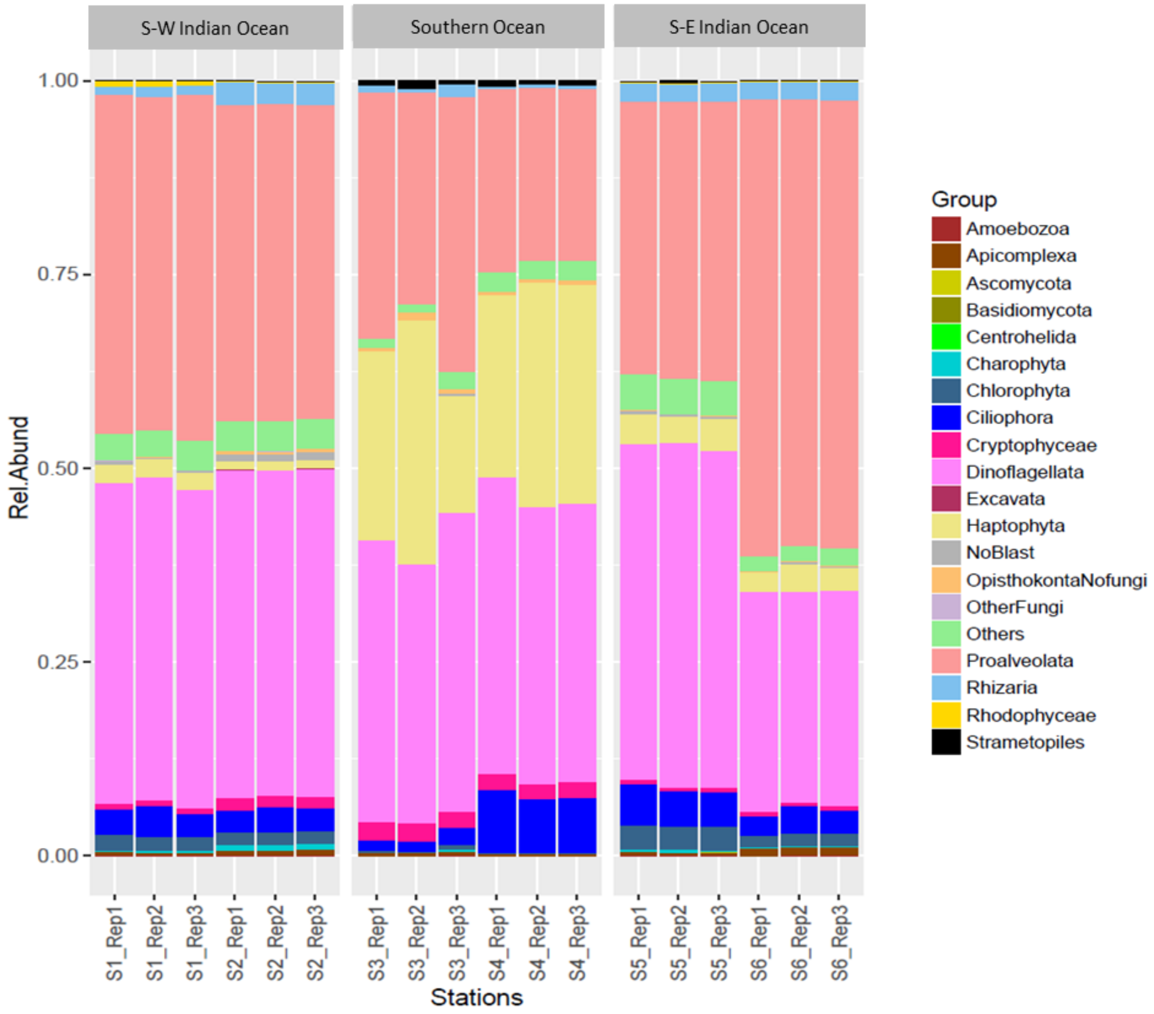


Figure 4.6: Eukaryotic relative abundance composition after removal of singletons (T1).

4.3.4 Geographic comparison of eukaryotic community

Eukaryotic OTU composition differed between the three main locations (Figure 4.6; PERMANOVA $F_{2,12} = 67.38$, $p = 0.001$*). Species richness was significantly different between the three locations (ANOVA $F_{2,12} = 30.22$, $p < 0.001$*). A post hoc Tukey's test showed that both the south-west and the south-east Indian Ocean were significantly different in the number of species to the Southern Ocean (p adj >0.0001), whilst the two southern Indian Ocean station were not significantly different (p adj = 0.851). The Bray-Curtis dissimilarity matrix analysed through hierarchical clustering showed how the six stations clustered as two different locations (above and below the APF) (Figure 4.7). Furthermore it can be observed the clustering of stations S2 and S5, suggesting some interchange across the southern Indian Ocean above the APF.

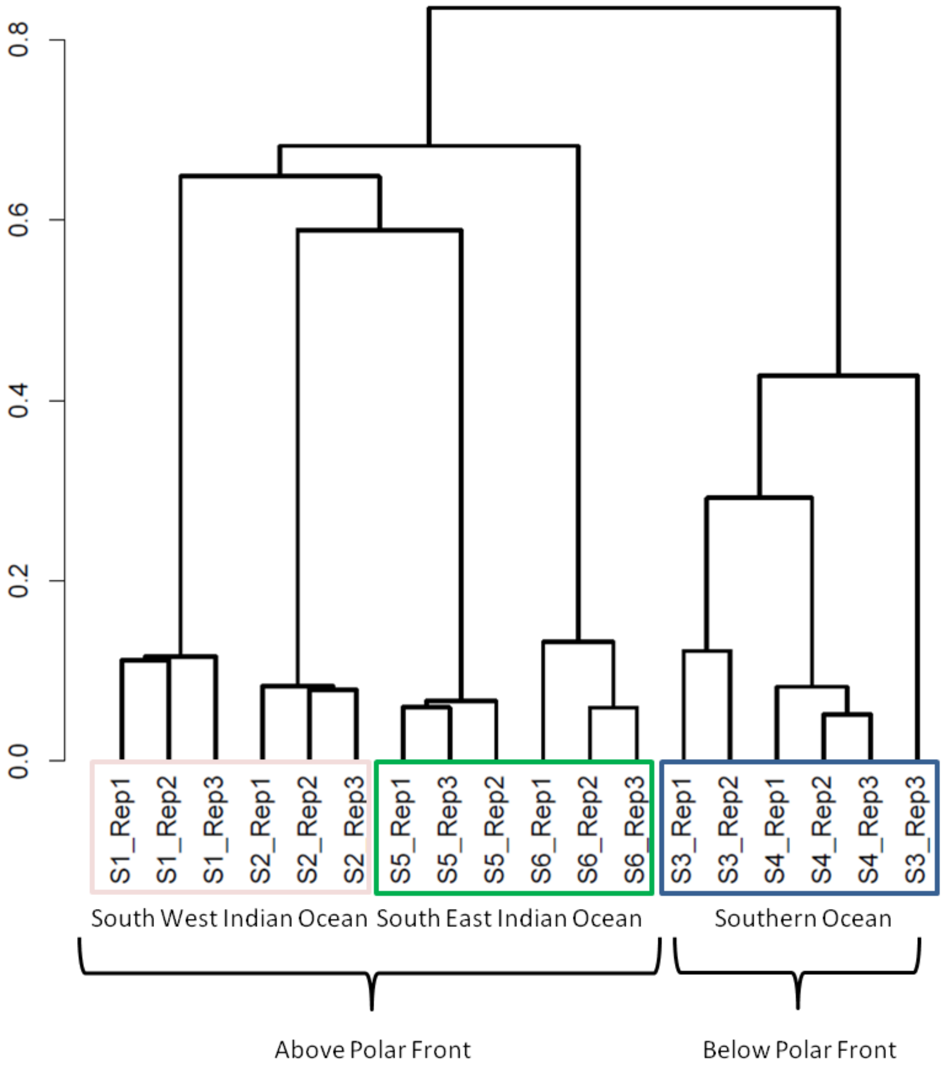


Figure 4.7: Hierarchical clustering based on Bray-Curtis dissimilarity matrix; eukaryotic dataset.

4.3.5 Biodiversity in the 0.45 µm filter permeate: viral contigs

The raw reads were assembled into a range between 5,000 contigs for station S1 and 35,000 for station S2 (Table 4.3). A selection of contigs was examined to confirm positive viral identification after annotation with the viral database; contigs were observed to have key viral features such as the presence of presumptive genes encoding viral tail components, major head proteins and viral capsid proteins (Supplementary Figure 4).

Viral sequences across the south Indian Ocean and the Southern Ocean were dominated by members of the order *Caudovirales* (60.57 ±5.96%); the lowest abundance (56%) for this order was observed in station S3, the maximum (71%) in station S2 (Figure 4.8, Supplementary Table 8). On average, members of the family *Myoviridae* represented 24.07 ±4.30%, members of the families *Siphoviridae* 21.39 ±3.32% and *Podoviridae* 13.92 ±5.21% of all the caudoviruses (Figure 4.8, Supplementary Table 8). Analysing caudoviruses separately, members of the family *Myoviridae* were the most abundant in four of the six stations representing on average 39.56 ±4.00% of this order (min 34.80% S4, max: 44.32% S2), while *Siphoviridae* members were the most abundant caudoviruses in both S4 and S6 (43% and 44%, respectively; Figure 4.9, Supplementary Table 8).

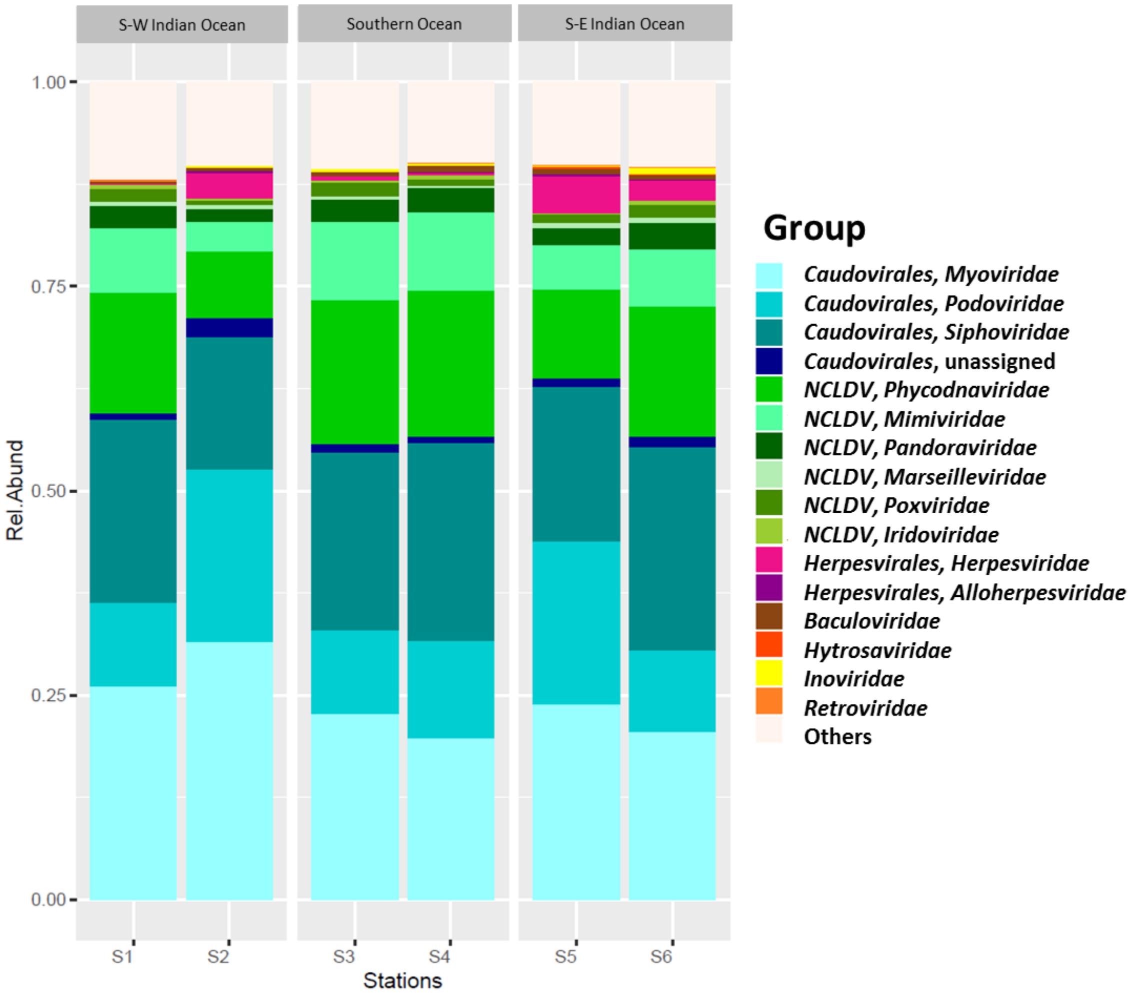


Figure 4.8: Viral contigs. Relative abundance of main viral group by station, separated by location.

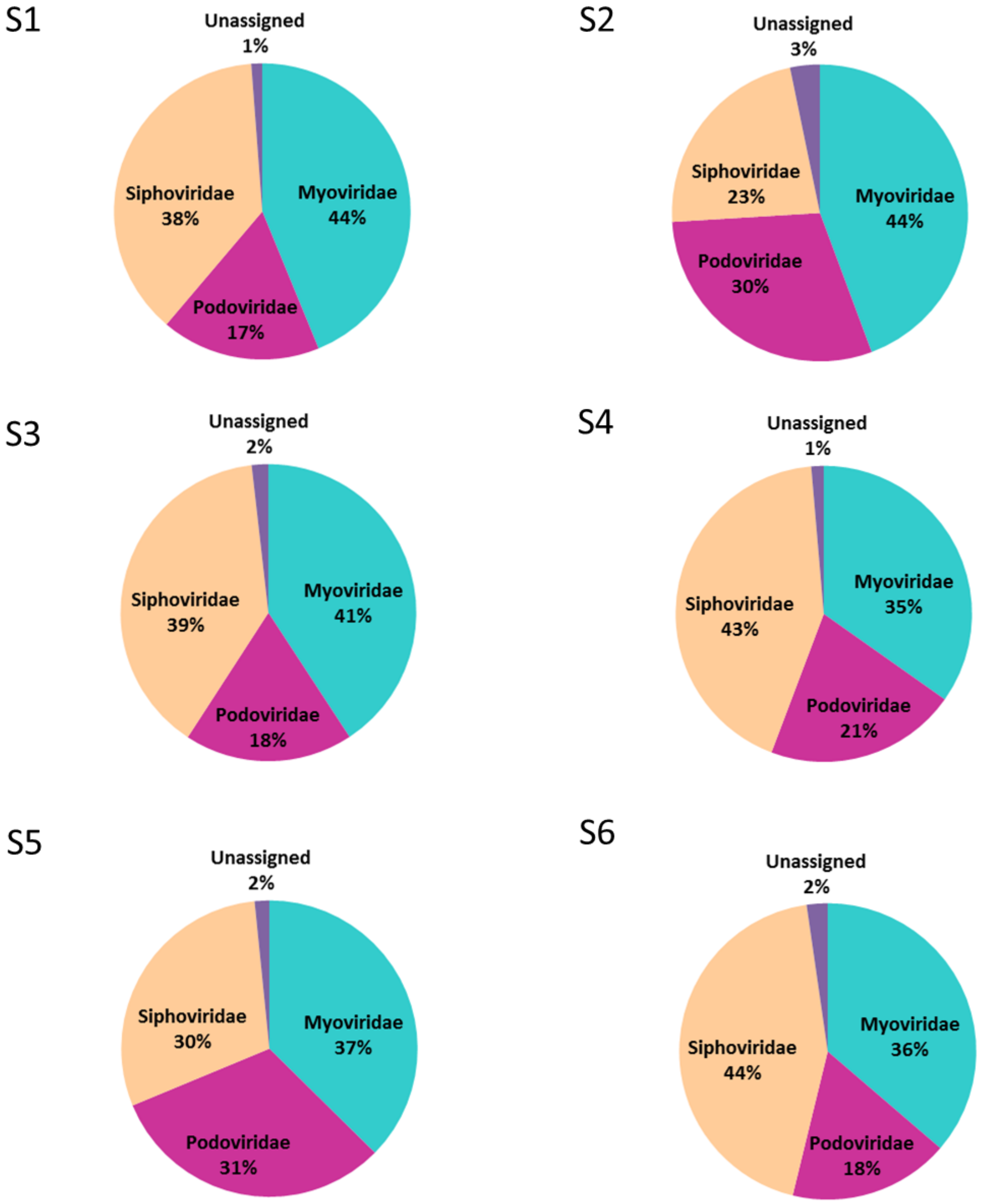


Figure 4.9: Order *Caudovirales*. Composition of the three caudovirus famililies (Myoviridae, Siphoviridae and Podoviridae) across the six stations.

Sequences assigned to the nucleocytoplasmic large DNA viruses (NCLDVs) represented the second most abundant group, comprising 26.35 ±6.85% of the contigs (Figure 4.8, Supplementary Table 8), with a minimum value in S2 (15.32%) and a maximum in S4 (32.67%). Phycodnaviruses represented half (49.67 ±2.41%) of the NCLDV sequences or 13.13 ±3.64% of all sequences (Figure 4.8, Supplementary Table 8). This viral grouping was dominated by generic chloroviruses in the south-west Indian Ocean, phaeoviruses in the Southern Ocean and by both chloroviruses and phaeoviruses in the south-east Indian Ocean (Figure 4.10). Specifically, phycodnavirus composition in the south-west Indian Ocean S1 and S2 differed in composition, with S1 characterised by the presence of chloroviruses and unassigned phycodnaviruses, which were present in the same numbers in S2; these were followed in abundance by prasino-, phaeo- and coccolithoviruses while S2 saw wider variation in the presence of phaeoviruses and prasinoviruses (Figure 4.10). Southern Ocean S3 and S4 were characterised by phaeoviruses, chloroviruses, unassigned phycodnavirus, prasino- and coccolithoviruses in order of abundance. In the south-east Indian Ocean phaeoviruses had similar presence as chloroviruses, followed by unassigned phycodnaviruses, prasino- and coccolithoviruses (Figure 4.10), while the highest numbers of coccolithovirus-like viruses were also observed here.

The NCLDV group was also characterised by the strong presence of members of the family *Mimiviridae* (26.86 ±2.56%, Figure 4.8), comprising on average 7.19 ±2.34% of the annotated contigs (min = 3.64% S2, max = 9.56% S3 and S4; Figure 4.8, Supplementary Table 8).

Viruses in the order *Herpesvirales* represented 1.97 ±1.92% of the annotated contigs, with a minimum (0.14%) in the south-west Indian Ocean S1 and a maximum (4.78%) in the south-east Indian Ocean S5 (Figure 4.8, Supplementary Table 8).

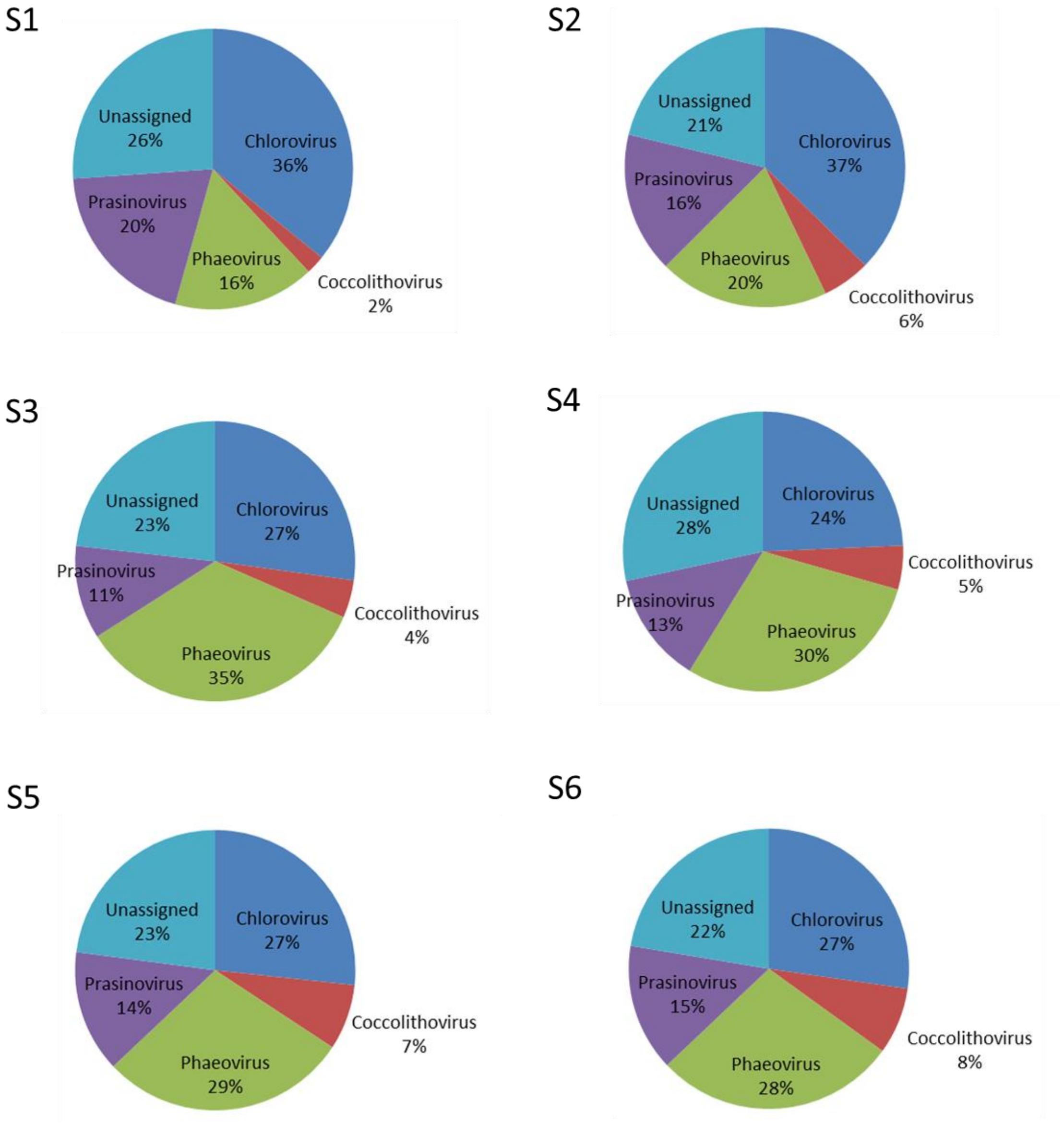


Figure 4.10: Phycodnavirus composition at the six stations.

4.3.6 Geographic comparison of the virome

Due to the nature of the metagenomic samples, i.e. the absence of replication, Log likelihood ratio statistics were used to test the null hypothesis of homogeneous viral distributions across all areas, and subsequently to test the influence of the polar front on viral dispersal. Results (Table 4.5) show that the polar front functioned as a barrier for viruses, with viral composition below the polar front being significantly different from stations located above the polar front (pchisq = 8.89E-120). The underlying viral distribution was also found to be significantly different between the three areas (Table 4.5).

Table 4.5: Pairwise log likelihood ratio statistics to test for differences in viral community composition between the three locations sampled in this study. Analyses were performed at Order level (*Caudovirales, Megavirales, Herpesvirales,* Other).

	S-W Indian Ocean	**S-W Indian Ocean**	**Southern Ocean**	**Above APF**
	:	**:**	**:**	**:**
	Southern Ocean	**S-E Indian Ocean**	**S-E Indian Ocean**	**Below APF**
H0.loglike	96484.7	97847.73	56698.81	125432.7
H1.loglike	96787.33	97901.93	56838.58	125709.7
lrs	605.2532	108.3967	108.3967	554.1192
pchisq	7.32E-131	2.43E-23	2.68E-60	8.89E-120

4.3.7 Comparison of the composition of the permeate (< 0.45µm) versus the cellular fraction (> 0.45µm)

Presence-absence analyses were performed across all fractions for each station to understand whether the metagenome contained unique OTUs due to the presence of eDNA. To do so the “genus” assignments from the ORFs on metagenomic assemble contigs were compared to the annotations from the amplicon sequences. For this analysis chloroplast and mitochondria OTUs present in the prokaryotic database were kept so as to verify overlap with the eukaryotic dataset. The majority of the sequences were not shared across the databases (Figure 4.11). A maximum of nine out of a possible 320 (0.57% of the overall dataset) eukaryotic genera could be detected in the permeate or eDNA fraction at one station (S2), while two stations (S1 and S3) shared no common sequences. A range of bacterial genera (9.58 to 15.25%) could however be found in common between the prokaryotic and eDNA databases (average 12.77 ±2.46%). Commonalities between the prokaryotic and eukaryotic database were due to the presence of chloroplast and mitochondria OTUs included in both databases (Table 4.6, Supplementary Figures 4 & 17-21).

The five most abundant genera identified within the eDNA permeate, made up almost half of the phylotypes detected (average = 52.57 ±9.37%, min = 41% in S2 and max = 63% in S3; Table 4.6). These genera were also found in the prokaryotic cellular fraction. Members of the genera *Alcanivorax* and *Marinobacter* were both found in three of the stations (S2, S5 and S6, Table 4.6).

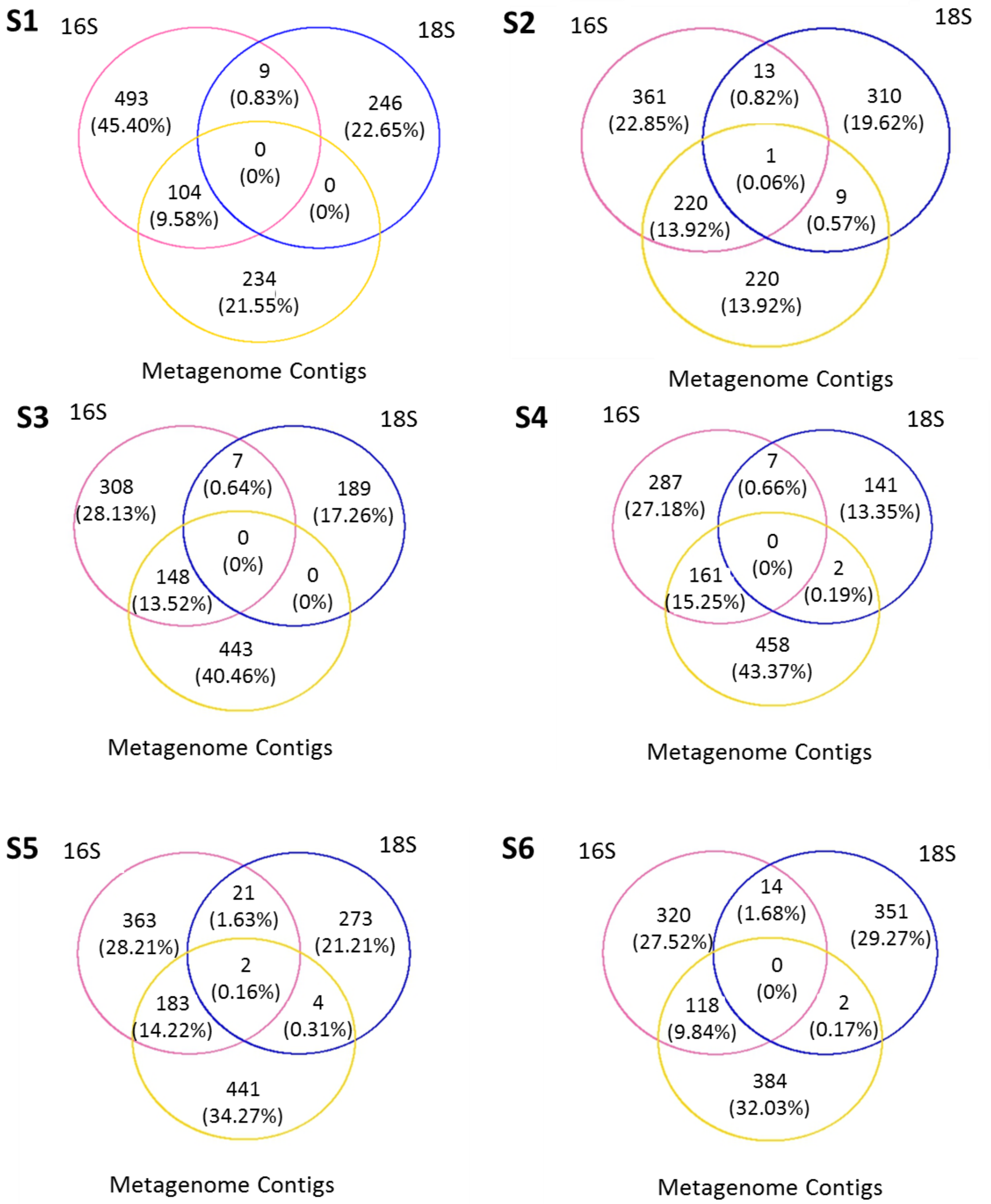


Figure 4.11: Presence-absence between the prokaryotes, eukaryotes and permeate. Comparisons were made on genus as the lowest level available, T1 for the prokaryotes and eukaryotes, and T0 on all contigs blasted with Refseq db.

Table 4.6: Five most abundant genera in the contigs annotated using Refseq db. Highlighted and in bold format, genera which were absent from the prokaryotic dataset.

S1	S2	S3	S4	S5	S6
Microbacterium 33%	*Halomonas 12%*	*Alteromonas 32%*	*Roseobacter 18%*	*Alteromonas 15%*	*Alcanivorax 16%*
Erythrobacter 24%	*Erythrobacter 10%*	*Sulfitobacter 10%*	*Sulfitobacter 12%*	*Marinobacter 11%*	***Oceanicola 12%***
Citomicrobium 4%	*Alcanivorax 7%*	*Halomonas 8%*	*Thalassolituus 7%*	***Oceanicola 8%***	*Methylophaga 9%*
Novosphingobium 2%	*Marinobacter 6%*	***Oceanibulbus 8%***	***Hypnomonas 5%***	*Thalassolituus 7%*	*Marinobacter 9%*
Arthrobacter 1%	*Methylophaga 6%*	*Erythrobacter 5%*	*Ruegeria 4%*	*Alcanivorax 7%*	***Hyphomonas 8%***
64%	41%	63%	46%	48%	54%

4.4 Discussion

Microbes are the biogeochemical engineers of life on Earth (Falkowski *et al.*, 2008), with bacterioplankton and phytoplankton contributing between 20 and 50% of the oceans' primary production (Cho and Azam, 1990; Azam *et al.*, 1983), as well as nearly half of the net Earth primary production (Field, 1998). The great abundance and importance of microbes in the oceans should drive our thinking towards fully understanding these microscopic organisms. It is only thanks to the relatively recent advent of NGS technologies and a modest number of global ocean expeditions (Sunagawa *et al.*, 2015; Rusch *et al.*, 2007; Kopf *et al.*, 2015; Gilbert *et al.*, 2011), that we are overcoming cultivability and sampling difficulties (Loman *et al.*, 2012a), and are starting to understand more about the world's most abundant inhabitants. Collection of all microbial components at the same time is still a major challenge due to the heterogeneity and patchiness of this system in time and space (Cao *et al.*, 2002; Deacon, 1982; Frederickson *et al.*, 2003; Seymour *et al.*, 2006; Boyd *et al.*, 2016), and is rendered even more challenging when multiple litres of water need to be processed for analysis. In this study, small volumes of water collected from CTD casts at the chlorophyll maximum were analysed in order to characterise the full range of protists to viruses from three different oceanographic systems: the south-east Indian Ocean, south-west Indian Ocean and the Southern Ocean, following the hypothesis that the APF works as an invisible physical barrier for microbial composition.

The comparative analyses of amplicon (prokaryotes and eukaryotes) and metagenomics (viruses and eDNA) datasets showed statistically significant differences of phylotype compositions between the two oceanic systems as well as across the three locations. Presence absence analyses showed that 30% and 44% of the eukaryotic and prokaryotic OTUs were shared among stations, leaving the majority of the sequences (70% of eukaryotic and 56% of prokaryotic OTUs) unique and present at only a specific station or location, thus reinforcing the hypothesis of spatial microbial patterns (Green and Bohannan, 2006), i.e. "the environment selects" rather than "everything in everyewhere".

The Bray-Curtis dissimilarity matrix analysed through hierarchical clustering for both amplicon datasets showed statistical differences above and below the APF, with the four south Indian Ocean samples sharing more in common with each other than the two samples collected in the Southern Ocean, especially across the Eukaryotic OTUs. Similarly, the log likelihood ratio statistic test conducted on the virome showed the presence of different communities above and below the front. It is therefore plausible to hypothesise that these differences can be attributed to the intense currents and thermocline of the APF, limiting the proliferation of specific groups of organisms due to the presence of extreme differences between the waters above and below the front. Differences in temperature, nutrients and minerals from the six stations exploited are characteristic of the South Indian Ocean and Southern Ocean (Donners and Drijfhout, 2004; Beal *et al.*, 2011; Popova *et al.*, 2000). These conditions provide a plausible explanation for the microbial community variation between these two systems.

During the International Census of Marine Microbes (ICoMM) 9.5 million DNA prokaryotic sequences clustered into 120,000 OTUs (Zinger *et al.*, 2011). The dataset presented in this study, prior to the removal of singleton, chloroplasts and mitochondria (T0) was characterised by 12.9 million prokaryotic sequences, which clustered into 133,500 OTUs. As seen in the ICoMM dataset (Zinger *et al.*, 2011), after the removal of singletons, chloroplast and mitochondria, almost half of the OTUs were still maintained. Differently to the ICoMM study, circa 70% of the sequences were retained for both the prokaryotes and eukaryotes, showing that the most abundant phylotypes were recovered.

The prokaryotic dataset showed dominance of Cyanobacteria in the water sampled from the south-west Indian Ocean, while the south-east Indian Ocean and the Southern Ocean were dominated by Proteobacteria. Specifically Gammaproteobacteria dominated both S5 and S6 from the south-east Indian Ocean and S3 from the Southern Ocean, whilst S4 was dominated by Alphaproteobacteria. These results are compatible with what was found during the Tara's global ocean sampling expedition, where Proteobacteria, specifically Alphaproteobacteria, dominated both surface waters and the deep chlorophyll maximum; Cyanobacteria and Gammaproteobacteria were the second most represented groups depending on location (Sunagawa *et al.*, 2015). Similar results were obtained during the ICoMM campaign in the surface open ocean with Alphaproteobacteria, Gammaproteobacteria, Cyanobacteria and Flavobacteria identified as the most abundant groups in the full datasets (Zinger *et al.*, 2011).

During the past few years, the Tara Ocean Expedition has contributed significantly to our understanding of microbes in the oceans; reporting for example dominance of Dinophyceae as OTU richness for the global pico-nanoplankton community, with almost 25,000 of the 87,000 annotated OTUs (28%) for the full eukaryotic dataset present in more than 40 of the 47 stations (de Vargas *et al.*, 2015). In contrast, the class Dinophyceae did not dominate the eukaryotic dataset but a similar ratio of protoalveolates and dinoflagellates was found. Specifically, protoalveolates dominated station S6 whilst dinoflagellates had higher concentrations in stations S5 and S4, and similar ratios of these microbes were found in stations S1, S2 and S3. Similar to the Tara's Ocean Expedition, the protoalveolates fraction was dominated by the Syndiniales groups I and II (de Vargas *et al.*, 2015) which were identified with previous nomenclature of MALV-I and MALV-II (Horiguchi, 2015). The two Southern Ocean stations (S3 and S4, sampled at the end of summer, March 2012) saw a higher abundance of haptophytes, due to presence of *Phaeocystis*. This relates to previous studies on the Southern Ocean, in which diatoms and haptophytes such as *Phaeocystis* were found more abundant in the more nutrient-rich polar fronts regions and continental shelves (Constable *et al.*, 2014). Furthermore, in the Ross sea *Phaeocystis* was the dominant primary producer in deeply mixed waters (20-50m) whereas diatoms dominated highly stratified waters (5-20m; (Arrigo *et al.*, 1999). Specifically Tara station 85 (sampled during summer, January 2011), based in the Southern Ocean, had a higher presence of haptophytes (de Vargas *et al.*, 2015).

Given the dominance of bacteria in the oceans, most marine viruses are assumed to be bacteriophages (Wommack and Colwell, 2000). Viruses, including bacteriophages, dominate oceanic waters, with approximately 10 million viruses per millilitre of seawater (Bergh *et al.*, 1989; Breitbart, 2012; Wilhelm and Suttle, 1999; Suttle, 2005). Metagenomic studies found that tailed viruses are the most abundant in the marine environment (Williamson *et al.*, 2008; Hurwitz and Sullivan, 2013; Williamson *et al.*, 2012; Chown *et al.*, 2015) and that myoviruses generally predominate, followed by podoviruses and then siphoviruses. However, it was reported that a hypersaline lagoon was dominated by siphoviruses, followed by podoviruses and then myoviruses (Williamson et al., 2008), showing that variation of this group might depend on abiotic conditions that affect the presence of its hosts. This understanding agrees with this study, where the annotated viral fraction was dominated by caudoviruses across all stations, with members of the family *Myoviridae* being most represented in S1, S2 and S5 whilst siphoviruses dominated in S4 and S6, and an equal ratio of Myo-Sipho ratio was observed in S3. NCLDVs, giant viruses infecting marine protists (Blanc-Mathieu and Ogata, 2016; Claverie and Abergel, 2013), were the second main viral group identified in the permeate, with familial phycodnaviruses representing almost half of this group in all six samples. This was previously assessed also for the Tara expedition, where just over half of the NCLDVs sequences were identified as phycodnaviruses with the other half identified as mimiviruses (Hingamp *et al.*, 2013), as also observed in this study. A higher presence of mimiviruses in the Southern Ocean samples (S3, S4) was observed and could be related to presence of Stramenopiles in these two stations as this relation has been previously hypothesised (Hingamp *et al.*, 2013). Furthermore, the presence of this family as second most common group of the NCLDVs reinforces the hypothesis that mimiviruses probably infect a wider variety of organisms than host-virus studies have indicated (Claverie *et al.*, 2009). Specifically for the family *Phycodnaviridae* it was expected, from previous studies, that there would be a primary presence of prasinoviruses (Hingamp *et al.*, 2013), whereas the dominance of chloroviruses was identified in S1 and S2, phaeoviruses in S5 and S6, and an equal ratio of both in S3 and S4.

Chloroviruses are known to infect and replicate in unicellular, chlorella-like green algae collected in freshwater (Dunigan *et al.*, 2006; Van Etten, 2003; Yamada *et al.*, 2006). They have also been reported to be able to replicate in humans and mice (Yolken *et al.*, 2014). Presence of chloroviruses in these marine samples allows speculation on the presence of alternative marine eukaryotic hosts. This is plausible as our knowledge of viruses infecting marine eukaryotes is still limited to only a few studies (Hingamp *et al.*, 2013), and biases in the isolation procedures against giant viruses are still commonplace (Van Etten, 2011). High abundance of dinoflagellates in the eukaryotic dataset suggests that these viruses could infect dinoflagellates as the most likely alternative hosts. Future studies targeting hosts and viruses could confirm this relation. Similarly, phaeoviruses are known to infect a broad range of brown macroalgae (Cock *et al.*, 2010), so the presence of this group of viruses in absence of their known hosts in the eukaryotic fraction could indicate an alternative host for this group as well, as hypothesised for mimiviruses (Claverie *et al.*, 2009).

Presence-absence analysis between the permeate and the cellular fraction collected on the filter showed that on average 13% of genera were identified in both the prokaryotic and the permeate datasets. The eukaryotic fraction on the other hand could not be described at all (0-0.57%). Four of the five most common prokaryotic genera identified in the permeate, representing nearly half of the permeate metagenomic dataset, could be found in the cellular amplicon dataset and therefore it could identify the presence of eDNA from a small proportion of the bacterial community. Interestingly, in three of the four south Indian Ocean stations *Alcanivorax* and *Marinobacter* were present. These organisms, known to degrade hydrocarbons

(Yakimov *et al.*, 1998; Moxley and Schmidt, 2012; Duran, 2010) could be a sign of oil-contaminated seawater due to active shipping routes.

The remaining genera, not identified in the filter but present in the permeate, could represent the presence of small bacteria passing through the 0.45µm filter (Hasegawa *et al.*, 2003; Tabor *et al.*, 1981; Anderson and Heffernan, 1965; Hahn, 2004), vesicles (Biller *et al.*, 2016) or "bacterial detritus" (Falkowski *et al.*, 2008). This identification was possible due to the sampling process for which singular organisms have not been isolated, which allowed the sampling of biodiversity otherwise not easily sampled (Biggs *et al.*, 2015; Bohmann *et al.*, 2014), identifying the permeate fraction as environmental DNA (eDNA). Finally, it is not possible to exclude that some of these "cellular" DNA could instead be of viral origin, since viral genes have been reported to match genes commonly found in the genomes of their prokaryotic and eukaryotic hosts (Wilson *et al.*, 2005; Baumann *et al.*, 2007; Filée *et al.*, 2007).

4.4.1 Conclusions

In this study it was shown that prokaryotic, eukaryotic and viral communities differ in composition in the south Indian Ocean and the Southern Ocean. These differences can be related to the open-ocean biological dispersal barrier created by the Antarctic Polar Front (Eastman, 1993; Thornhill *et al.*, 2008). Variations in community composition were observed also on the south-west and south-east Indian Ocean, with the prokaryotic community being more separated than the eukaryotes. Differences in the host fraction were reflected into the viral composition across the three sampling location. Furthermore the increase in haptphytes in the Southern Ocean was reflective of an increase of large eukaryotic viruses.

These differences, affecting the microbial communities, can be attributed to the location of the south-east samples collected below the Subtropical front (Balch *et al.*, 2016). As found in the Tara ocean global expedition study (Sunagawa *et al.*, 2015), this study indicates that water temperature, which is a major defining characteristic of the different stations above and below the APF, plays an important role for determining microbial community dispersal. Finally, results showed in this study unequivocally demonstrates that the composition of the cellular amplicon fraction differs dramatically from the eDNA permeate as sampled; therefore raising the efficacy of the eDNA being used to monitor aquatic biodiversity.

Chapter 5: General Conclusions

The Pew Ocean Commission (www.pewoceans.org/oceans/ oceans_report.asp) occurred in 2003 and highlighted rising concerns for the health and biodiversity of the oceans showing the necessity for a microbial exploration of seas and oceans (Azam and Worden, 2004). If the subsequent increase in global ocean microbial surveys on one side has exposed the physiological and biogeochemical functions of marine microbes, on the other side they have revealed a gap in the understanding of microbes ecological niches. Further advancements in modern cultivation-independent tools and new cultivation technologies (Loman *et al.*, 2012b) are proving to be effective in improving the characterisation of marine microbes. These advancements have allowed the diversification of sequencing platforms utilised for microbial studies. If the 454 pyrosequencing technology was at the base of early microbial studies, recent reductions in the cost of Illumina technologies have favoured the use of this platform to study microbial population (Caporaso *et al.*, 2011). The wide use of the Illumina platforms in both terrestrial and marine environments (Gilbert *et al.*, 2014; Caporaso *et al.*, 2011, 2012b; Gilbert *et al.*, 2012) together with the lower costs, have made it the method of choice for this project. The same water collected from a CTD was used for two different sequencing methods: amplicon sequencing was chosen for the characterisation of both prokaryotes and eukaryotes using the 16S and 18S rRNA genes respectively whilst, due to lack of conserved viral genes (Breitbart *et al.*, 2002, 2004), metagenomic shotgun sequencing was performed to look at viral composition. The use of metagenomic shotgun sequencing also allowed the same sample to be used to explore the residual environmental genetic material (eDNA) present in the water sampled. This represents an important and innovative approach, since few studies have looked at the whole community (Zinger *et al.*, 2012) following limitations of previous methods. This study demonstrates that an alternative protocol, which allows for characterisation of the most abundant phylotypes using only a small volume of water, can reduce the sampling effort required to describe microbial diversity. Reduction in sampling effort makes more practical, compared to the traditional large volume cruise based programmes, the gathering of a greater number of samples within a region, or sample stations over annual cycles, as well as collecting and preserving a larger number of samples for future analysis if needed.

A number of tests were utilised to provide evidence as to the robustness of the method prior to the description of the microbial community, as well as some indirect comparisons could be made with some previous studies (Zinger *et al.*, 2011; Williamson *et al.*, 2012; de Vargas *et al.*, 2015; Sunagawa *et al.*, 2015). Replication and saturation were addressed to test the suitability for its use on the prokaryotic dataset; to the six southern hemisphere samples were added two harmful algal bloom samples and a northern hemisphere sample to increase diversity of environments. The use of PCR replicates allowed the removal of erroneous sequences whilst maintaining rare phylotypes, adding confidence that only genuine OTUs were kept whilst sequencing errors and artifacts were removed. This study proves that the removal of singletons is an essential step and the failure to remove these sequences from the dataset would lead to the inclusion of errors in the analysis. Despite the reduction in the number of OTUs, the number of reads was not significantly affected, demonstrating that the OTUs removed were most likely representative of errors and artifacts, and the reads removed represented less than 1% of the total dataset.

Throughout this study, confidence was built towards the use of a small volume of water, 250ml, which allowed the further characterisation of the microbial fraction from the stations analysed, as well as the indirect comparison with sampling stations located in a similar area to this study. These showed a comparable characterisation of the most abundant phylotypes between the studies. Therefore the use of replication and different filter or thresholds were further applied to both amplicon datasets, prokaryotic and eukaryotic, to better discern true and rare phylotypes over background sequencing noise (errors, artifacts and contamination). Following the removal of singletons, four extra filters were tested, but the application of these more stringent thresholds resulted in the elimination from the dataset of sequences belonging to genuine rare microbiota. Therefore these more stringent filters were not used in the final analysis of study samples.

Together with the description of the microbial diversity, an important factor in the understanding of microbial communities is their movement across different systems. In the marine environment, the presence of microbial biogeographic patterns is still an open debate (Martiny *et al.*, 2006; Staley, 1997; Finlay, 2002). Therefore the identification of the presence of marine barriers, affecting genetic flow, could lead to a better understanding of these patterns. Consistency in time and space of the sampling method utilised, together with confidence in the dataset processing, allowed the study to test the hypothesis that marine barriers can affect microbial composition. To do so this study looked at pairs of sampling stations in three geographic locations based on both sides of the Antarctic Polar Front (APF). Intense currents such as those of the APF (Eastman, 1993), that are known to affect the distribution of some eukaryotes (Thornhill *et al.*, 2008; Shaw *et al.*, 2004; Hunter and Halanych, 2008), present a good example to test the effects of these currents as a barrier for microbial genetic flow. Therefore this study represents a step forward by analysing the composition of the most abundant microbial phylotypes on these sampling stations, located on both sides of the APF. The statistically significant differences in distribution of prokaryotes, eukaryotes and viruses in the three locations sampled, showed that strong fronts such as the APF can affect microbial communities' composition. Specifically it was possible to observe variation in the most abundant phylotypes on the two sides of the front. For the prokaryotic community significant differences were present across all three locations examined, with Cyanobacteria dominating the south-east Indian Ocean, Gammaproteobacteria the south-west Indian Ocean and a combination of Alpha- and Gammaproteobacteria in the Southern Ocean. This variation in composition was also observed, in previous studies, for different marine environments (Sunagawa *et al.*, 2015; Zinger *et al.*, 2011) where the distribution of Cyanobacteria, Alpha- and Gammaproteobacteria differed depending on the sampled site. Similar observations could be made for the eukaryotes, despite Alveolates dominating the eukaryotic fraction, within this group differences in the ratio of Protoalveolata and Dinoflagellata were observed together with a considerable increase in the haptophyte community in the Southern Ocean. The presence of higher numbers of haptophytes in the Southern Ocean was expected as it had been previously described for this oceanic system (Arrigo *et al.*, 1999; de Vargas *et al.*, 2015); nevertheless it wasn't possible to observe the dominance of Dinophyceae within this dataset as contrarily found previously from various locations (de Vargas *et al.*, 2015). Differences in environmental conditions, such as nutrients and temperature, between the three locations could be the leading cause of shifts in community composition; the environment selects. However, it also a symptom of the effects of the APF which exclusively separates the two oceans (Indian and Southern Ocean) creating specific environments in which the most adaptable organisms will prevail and therefore are found in a greater abundance.

If abiotic factors are responsible for variations of the host community, then consequently the viral fraction will be affected. In the six stations charaterised by this study, likewise the hosts, the viral fraction showed differences in its distribution. Within the order *Caudovirales*, fluctuations in the ratios of the three families were observed stationwise and mainly across locations as observed for sampling stations by Williamson *et al.* (2008). Furthermore, despite the dominance of caudoviruses, it was interesting to observe the increase of NCLDVs in correlation with an increase in the haptophytes in the Southern Ocean. NCLDVs knowledge is still scarce and often overlooked (Monier *et al.*, 2008) despite being identified both in this study, and in previous studies, as the second most abundant viral group (Hingamp *et al.*, 2013). Of these large eukaryotic viruses two familes prevail, *Phycodnaviridae* and *Mimiviridae*. Interestingly within the family *Phycodnaviridae* there was no observed predominance of prasinovirus, as observed in the Tara Ocean dataset (Hingamp *et al.* 2013), and furthermore there was a higher presence of chloroviruses and phaeoviruses, which are known to infect respectively freshwater (Dunigan *et al.*, 2006; Van Etten, 2003) and brown algae (Cock *et al.*, 2010). It has been observed that NCLDVs such as mimivirus can infect a wide variety of organisms (Claverie *et al.*, 2009) and it has been reported that choroviruses are able to replicate in mammals (Yolken *et al.*, 2014). It is therefore possible that both the chloroviruses and phaeovirses identified in this study represent similar viruses to the one infecting freshwater green algae and brown algae, respectively, but are infecting alternative hosts within the marine environment.

Reports are showing that dinoflagellates, known to be infected by RNA viruses (Tomaru *et al.*, 2009), can be infected by NCLDVs (Nagasaki *et al.*, 2006; Nagasaki, 2008). Due to the abundance of haptophytes and NCLDVs, without another identified host, it can be speculated that some of these giant viruses could be infecting these specific organisms. Future work, including additional network analsyis, will help unveil and understand hidden relationships between these large viruses and possible hosts as is happening for small RNA viruses (Steward *et al.*, 2013).

Whilst the results for amplicon and metagenomic analysis proved their usefulness as a tool for understanding the microbial community, this study demonstrates that caution must be used when drawing conclusions based on eDNA when tracking rare or endangered species. Previously it has been proposed that eDNA could be used as a monitoring tool (Valentini *et al.*, 2016) to determine whether an invasion has taken place (Dejean *et al.*, 2012) or to track an endangered species (Ikeda *et al.*, 2016). Results from the analysis of the six stations unequivocally showed that the composition of the cellular fraction differed dramatically from the DNA contained in the environmental fraction. While it was possible to identify a small portion (~10%) of bacteria both in the cellular amplicon and the environmental metagenomic fractions, it was impossible to detect eukaryotic DNA in the permeate, therefore showing that eDNA is not a feasible tool to monitor eukaryotic diversity or the presence of rare species. This highlights the necessity of more in depth studies to understand the role of eDNA, and its suitability for detecting the passage of eukaryotic organisms in the water fraction, and that previous conclusions based purely on eDNA sampling must be treated with caution.

To conclude, this method for the identification of the most abundant phylotypes from a single small volume of water provides a powerful monitoring tool, which allows the clear documentation of shifts in populations on both a local and oceanic scale. The future inclusion of time series to monitor the microbial composition will help answering questions on the composition of the marine microbiome, due to seasonal variation or stress factors such as the increase in temperature, pH or changes in salinity. This will provide robust data points that could help, not only the progress of microbial research but consequently providing a reference to assess water quality. The application of this method for the monitoring of specific communities of economic importance, for example around open ocean aquaculture farming areas, can rapidly highlight variation in the microbial community structure and allow a prompt response. In the case of harmful algal blooms, sewage leaks, or oil spills a rapid detection of a microbial community shift, based upon long term monitoring, provides the means to investigate further and predict potential long term effects. Furthermore it will help keeping track of changes in the dominant populations due to climate change and ocean acidification, and therefore provide a practical tool to better understand the complex role of marine microbes in the environment.

Ecologically, this study has shown for the first time that the Antarctic Polar Front can create a genetic barrier for microbial genetic flow. This frontal system, characterised by intense currents and thermoclines (Eastman, 1993; Thornhill et al., 2008), can create different abiotic conditions in the oceans that it separates. This invisible barrier will affect the composition of the microbial communities and consequently their viruses, with distinct host-virus interactions unique to either side of the divide. It can be hypothesised that the study of other frontal systems will demonstrate the presence of additional barriers in microbial dispersal, and used to predict the effects on microbial distribution with predicted changes in these currents due to global climate change (Solomon *et al.*, 2007). It was exciting to observe the increase in the eukaryotic viruses in the Antarctic system, as well as the presence of viruses known to infect different hosts, or previously only known to be present in non-marine environments. Only by sampling the total microbial community, including viruses and their hosts, alongside oceanographic studies we can characterise and understand the essential role of these invisible enitities. This study provides a practicable tool for doing so and opens the door for future discoveries.

Reference

Ackermann H-W. (2003). Bacteriophage observations and evolution. *Res Microbiol* **154**: 245–251.

Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, *et al.* (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol* **52**: 399–451.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–10.

Amaral-Zettler L, Artigas LF, Baross J, Bharathi P.A. L, Boetius A, Chandramohan D, *et al.* (2010). A Global Census of Marine Microbes. In: *Life in the World's Oceans*. Wiley-Blackwell: Oxford, UK, pp 221–245.

Anderson JIW, Heffernan WP. (1965). Filterable Marine Bacteria Isolation and Characterization of Filterable Marine Bacteria '. *J Bacteriol* **90**: 1713–1718.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards R a., Carlson C, *et al.* (2006). The marine viromes of four oceanic regions. *PLoS Biol* **4**: 2121–2131.

Armstrong E, Yan L, Boyd KG, Wright PC, Burgess JG. (2001). The symbiotic role of marine microbes on living surfaces. *Hydrobiologia* **461**: 37–40.

Arrigo, Robinson, Worthen, Dunbar, DiTullio, VanWoert, *et al.* (1999). Phytoplankton community structure and the drawdown of nutrients and CO2 in the southern ocean. *Science* **283**: 365–7.

Azam F, Fenchel T, Field J, Gray J, Meyer-Reil L, Thingstad F. (1983). The Ecological Role of Water-Column Microbes in the Sea. *Mar Ecol Prog Ser* **10**: 257–263.

Azam F, Worden AZ. (2004). Microbes, Molecules, and Marine Ecosystems. *Science (80-)* **303**: 1622–1624.

Balch WM, Bates NR, Lam PJ, Twining BS, Rosengard SZ, Bowler BC, *et al.* (2016). Factors regulating the Great Calcite Belt in the Southern Ocean and its biogeochemical significance. *Global Biogeochem Cycles* **30**: 1124–1144.

Baldauf SL. (2003). The deep roots of eukaryotes. *Science* **300**: 1703.

Barton AD, Irwin AJ, Finkel Z V., Stock CA. (2016). Anthropogenic climate change drives shift and shuffle in North Atlantic phytoplankton communities. *Proc Natl Acad Sci* **113**: 2964–2969.

Baumann S, Sander A, Gurnon JR, Yanai-Balser GM, Van Etten JL, Piotrowski M. (2007). Chlorella viruses contain genes encoding a complete polyamine biosynthetic pathway. *Virology* **360**: 209–217.

Beal LM, De Ruijter WPM, Biastoch A, Zahn R. (2011). On the role of the Agulhas system in ocean circulation and climate. *Nature* **472**: 429–36.

Becking LB. (1934). Geobiologie of inleiding tot de milieukunde. W.P. Van Stockum & Zoon: Den Haag.

Behnke A, Engel M, Christen R, Nebel M, Klein RR, Stoeck T. (2011). Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environ Microbiol* **13**: 340–349.

Beijerinck MW. (1913). De infusies en de ontdekking der backteriën. Jaarboek van de Koninklijke

Akademie van Wetenschappen. Müller: Amsterdam, The Netherlands. In: *Müller: Amsterdam, The Netherlands,*. pp 1–28.

Belkin IM, Gordon AL. (1996). Southern Ocean fronts from the Greenwich meridian to Tasmania. *J Geophys Res* **101**: 3675.

Bell EA, Boehnke P, Harrison TM, Mao WL. (2015). Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proc Natl Acad Sci U S A* **112**: 14518–21.

Bergey DH (David H, Krieg NR, Holt JG. (1984). Bergey's manual of systematic bacteriology. Williams & Wilkins: Baltimore MD.

Bergh O, Børsheim KY, Bratbak G, Heldal M. (1989). High abundance of viruses found in aquatic environments. *Nature* **340**: 467–8.

Biggs J, Ewald N, Valentini A, Gaboriaud C, Dejean T, Griffiths RA, *et al.* (2015). Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (Triturus cristatus). *Biol Conserv* **183**: 19–28.

Biller SJ, Schubotz F, Roggensack SE, Thompson AW, Summons RE, Chisholm SW. (2016). Bacterial Vesicles in Marine Ecosystems Accessed. *Science (80-)* **343**: 183–186.

Blanc-Mathieu R, Ogata H. (2016). DNA repair genes in the Megavirales pangenome. *Curr Opin Microbiol* **31**: 94–100.

Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, *et al.* (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends Ecol Evol* **29**: 358–367.

Bonnain C, Breitbart M, Buck KN. (2016). The Ferrojan Horse Hypothesis: Iron-Virus Interactions in the Ocean. *Front Mar Sci* **3**: 1–11.

Boone DR, Castenholz RW, Garrity GM. (2001). Bergey's manual of systematic bacteriology. Springer.

Bork P, Bowler C, de Vargas C, Gorsky G, Karsenti E, Wincker P. (2015). Tara Oceans studies plankton at planetary scale. *Science (80-)* **348**: 873–873.

Boyd PW, Cornwall CE, Davison A, Doney SC, Fourquez M, Hurd CL, *et al.* (2016). Biological responses to environmental heterogeneity under future ocean conditions. *Glob Chang Biol* 2633–2650.

Breitbart M. (2012). Marine viruses: truth or dare. *Ann Rev Mar Sci* **4**: 425–48.

Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P, *et al.* (2004). Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci* **271**: 565–574.

Breitbart M, Rohwer F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13**: 278–84.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, *et al.* (2002). Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**: 14250–5.

Breitbart M, Thompson L, Suttle C, Sullivan M. (2007). Exploring the Vast Diversity of Marine Viruses. *Oceanography* **20**: 135–139.

Brown M V, Philip GK, Bunge JA, Smith MC, Bissett A, Lauro FM, *et al.* (2009). Microbial community structure in the North Pacific ocean. *ISME J* **3**: 1374–1386.

Brum JR, Culley AI, Steward GF. (2013a). Assembly of a marine viral metagenome after physical fractionation. *PLoS One* **8**: e60604.

Brum JR, Hurwitz BL, Schofield O, Ducklow HW, Sullivan MB. (2015a). Seasonal time bombs: dominant temperate viruses affect Southern Ocean microbial dynamics. *ISME J* 1–13.

Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, *et al.* (2015b). Patterns and ecological drivers of ocean viral communities. *Science (80-)* **348**: 1261498–1261498.

Brum JR, Schenck RO, Sullivan MB. (2013b). Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J* **7**: 1738–1751.

Brun P, Kiørboe T, Licandro P, Payne MR. (2016). The predictive skill of species distribution models for plankton in a changing climate. *Glob Chang Biol* **22**(9):3170-81.

Buchan A, LeCleir GR, Gulvik CA, González JM. (2014). Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat Rev Microbiol* **12**: 686–698.

Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brüssow H. (2003). Phage as agents of lateral gene transfer. *Curr Opin Microbiol* **6**: 417–424.

Cao Y, Williams DD, Larsen DP. (2002). Comparison of ecological communities: the problem of sample representatoveness. *Ecol Monogr* **72**: 41–56.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.

Caporaso JG, Lauber CL, Walters W a, Berg-Lyons D, Huntley J, Fierer N, *et al.* (2012a). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**: 1621–4.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, *et al.* (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* **108 Suppl**: 4516–22.

Caporaso JG, Paszkiewicz K, Field D, Knight R, Gilbert JA. (2012b). The Western English Channel contains a persistent microbial seed bank. *ISME J* **6**: 1089–93.

Cheung MK, Au CH, Chu KH, Kwan HS, Wong CK. (2010). Composition and genetic diversity of picoeukaryotes in subtropical coastal waters as revealed by 454 pyrosequencing. *ISME J* **4**: 1053–1059.

Chicote E, García AM, Moreno DA, Sarró MI, Lorenzo PI, Montero F. (2005). Isolation and identification of bacteria from spent nuclear fuel pools. *J Ind Microbiol Biotechnol* **32**: 155–162.

Cho B, Azam F. (1990). Biogeochemical significance of bacterial biomass in the ocean’s euphotic zone. *Mar Ecol Prog Ser* **63**: 253–259.

Chow CET, Suttle C a. (2015). Biogeography of Viruses in the Sea. *Annu Rev Virol* **2**: 41–66.

Chow CET, Winget DM, White R a., Hallam SJ, Suttle C a. (2015). Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. *Front Microbiol* **6**: 1–15.

Chown SL, Clarke A, Fraser CI, Cary SC, Moon KL, McGeoch MA. (2015). The changing form of Antarctic biodiversity. *Nature* **522**: 431–438.

Claverie J. (2006). Viruses take center stage in cellular evolution. *Genome Biol* **7**: 110.

Claverie J, Abergel C. (2013). Open questions about giant viruses. *Adv Virus Res* **85**: 25–56.

Claverie JM, Grzela R, Lartigue A, Bernadac A, Nitsche S, Vacelet J, *et al.* (2009). Mimivirus and

Mimiviridae: Giant viruses with an increasing number of potential hosts, including corals and sponges. *J Invertebr Pathol* **101**: 172–180.

Claverie JM, Ogata H, Audic S, Abergel C, Suhre K, Fournier PE. (2006). Mimivirus and the emerging concept of ‘giant’ virus. *Virus Res* **117**: 133–144.

Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, *et al.* (2010). The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* **465**: 617–621.

Colson P, de Lamballerie X, Fournous G, Raoult D. (2012). Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. *Intervirology* **55**: 321–32.

Colson P, De Lamballerie X, Yutin N, Asgari S, Bigot Y, Bideshi DK, *et al.* (2013). ‘Megavirales’, a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Arch Virol* **158**: 2517–21.

Constable AJ, Melbourne-Thomas J, Corney SP, Arrigo KR, Barbraud C, Barnes DKA, *et al.* (2014). Climate change and Southern Ocean ecosystems I: How changes in physical habitats directly affect marine biota. *Global Change Biology*. **20**(10):3004-25.

Correa AMS, Welsh RM, Vega Thurber RL. (2013). Unique nucleocytoplasmic dsDNA and +ssRNA viruses are associated with the dinoflagellate endosymbionts of corals. *ISME J* **7**: 13–27.

Culley AI. (2011). Virophages to viromes: A report from the frontier of viral oceanography. *Curr Opin Virol* **1**: 52–57.

Curtis TP, Sloan WT, Scannell JW. (2002). Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci* **99**: 10494–10499.

Dalrymple GB. (2001). The age of the Earth in the twentieth century: a problem (mostly) solved. *Geol Soc London, Spec Publ* **190**: 205–221.

Deacon GER. (1982). Physical and biological zonation in the Southern Ocean. *Deep Sea Res Part A Oceanogr Res Pap* **29**: 1–15.

Dejean T, Valentini A, Miquel C, Taberlet P, Bellemain E, Miaud C. (2012). Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog Lithobates catesbeianus. *J Appl Ecol* **49**: 953–959.

Delaroque N, Boland W. (2008). The genome of the brown alga Ectocarpus siliculosus contains a series of viral DNA pieces, suggesting an ancient association with large dsDNA viruses. *BMC Evol Biol* **8**: 110.

Dìez B, Pedrós-alió C, Massana R. (2001). Study of Genetic Diversity of Eukaryotic Picoplankton in Different Oceanic Regions by Small-Subunit rRNA Gene Cloning and Sequencing Study of Genetic Diversity of Eukaryotic Picoplankton in Different Oceanic Regions by Small-Subunit rRNA Gene Cloning and. *Appl Environ Microbiol* **67**: 2932–2941.

Donlon CJ, Martin M, Stark J, Roberts-Jones J, Fiedler E, Wimmer W. (2012). The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sens Environ* **116**: 140–158.

Donners J, Drijfhout SS. (2004). The Lagrangian View of South Atlantic Interocean Exchange in a Global Ocean Model Compared with Inverse Model Results. *J Phys Oceanogr* **34**: 1019–1035.

Dorigo U, Fontvieille D, Humbert JF. (2006). Spatial variability in the abundance and composition of the free-living bacterioplankton community in the pelagic zone of Lake Bourget (France). *FEMS Microbiol Ecol* **58**: 109–119.

Ducklow HW, Steinberg DK, Buesseler KO. (2001). Upper ocean carbon export and the biological pump. *Oceanography* **14**: 50–58.

Dunigan DD, Fitzgerald LA, Van Etten JL. (2006). Phycodnaviruses: A peek at genetic diversity. *Virus Res* **117**: 119–132.

Duran R. (2010). Marinobacter. In: Timmis KN (ed) Vol. 1. *Handbook of Hydrocarbon and Lipid Microbiology*. Springer Berlin Heidelberg: Berlin, Heidelberg, pp 1725–1735.

Eastman JT. (1993). Antarctic fish biology : evolution in a unique environment. Academic Press.

Eberlein T, Van de Waal D, Brandenburg K, John U, Voss M, Achterberg E, *et al.* (2016). Interactive effects of ocean acidification and nitrogen limitation on two bloom-forming dinoflagellate species. *Mar Ecol Prog Ser* **543**: 127–140.

Edwards RA, Rohwer F. (2005). Viral metagenomics. *Nat Rev Microbiol* **3**: 504–10.

Van Etten JL. (2011). Another really, really big virus. *Viruses* **3**: 32–46.

Van Etten JL. (2003). Unusual Life Style of Giant Chlorella Viruses. *Annu Rev Genet* **37**: 153–195.

Falkowski PG, Fenchel T, Delong EF. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science (80-)* **320**: 1034–1039.

Faust K, Raes J. (2012). Microbial interactions: from networks to models. *Nat Rev Microbiol* **10**: 538–550.

Feil EJ. (2004). Small change: keeping pace with microevolution. *Nat Rev Microbiol* **2**: 483–495.

Field CB. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science (80-)* **281**: 237–240.

Filée J. (2013). Route of NCLDV evolution: the genomic accordion. *Curr Opin Virol* **3**: 595–9.

Filée J, Siguier P, Chandler M. (2007). I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. *Trends Genet* **23**: 10–15.

Finlay BJ. (2002). Global Dispersal of Free-Living Microbial Eukaryote Species. *Science (80-)* **296**: 1061–1063.

Flombaum P, Gallegos JL, Gordillo R a, Rincón J, Zabala LL, Jiao N, *et al.* (2013). Present and future global distributions of the marine Cyanobacteria Prochlrococcus and Synechococcus. *Pnas* **110**: 9824–9829.

Follows MJ, Dutkiewicz S. (2011). Modeling diverse communities of marine microbes. *Ann Rev Mar Sci* **3**: 427–451.

Follows MJ, Dutkiewicz S, Grant S, Chisholm SW. (2007). Emergent Biogeography of Microbial Communities in a Model Ocean. *Science (80-)* **315**: 1843–1846.

Forterre P. (2010). Defining Life: The Virus Viewpoint. *Orig Life Evol Biosph* **40**: 151–160.

Forterre P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res* **117**: 5–16.

Frederickson CM, Short SM, Suttle CA. (2003). The Physical Environment Affects Cyanophage Communities in British Columbia Inlets. *Microb Ecol* **46**: 348–357.

Fuhrman JA. (1992). Bacterioplankton role in cycling of organic matter: The microbial food web. *Prim Prod Biogeochem Cycles Sea* **43**: 361–379.

Fuhrman JA. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541–8.

Fuhrman JA. (2009). Microbial community structure and its functional implications. *Nature* **459**: 193–9.

Fuhrman JA, Noble RT. (1995). Viruses and protists cause similar bacterial mortality in coastal seawater. *Limnol Oceanogr* **40**: 1236–1242.

Ghiglione J, Larcher M, Lebaron P. (2005). Spatial and temporal scales of variation in bacterioplankton community structure in the NW Mediterranean Sea. *Aquat Microb Ecol* **40**: 229–240.

Gilbert JA, Bailey M, Field D, Fierer N, Fuhrman JA, Hu B, *et al.* (2011). The Earth Microbiome Project: The Meeting Report for the 1st International Earth Microbiome Project Conference, Shenzhen, China, June 13th-15th 2011. *Stand Genomic Sci* **5**: 243–247.

Gilbert JA, Jansson JK, Knight R. (2014). The Earth Microbiome project: successes and aspirations. *BMC Biol* **12**: 69.

Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B, *et al.* (2012). Defining seasonal marine microbial community dynamics. *ISME J* **6**: 298–308.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**: 60–63.

Glansdorff N, Xu Y, Labedan B. (2008). The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol Direct*. **9:** 3:29

Glenn TC. (2011). Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11**: 759–69.

Gordon AL. (1986). Interocean exchange of thermocline water. *J Geophys Res* **91**: 5037–5046.

Green J, Bohannan BJM. (2006). Spatial scaling of microbial biodiversity. *Trends Ecol Evol* **21**: 501–507.

Goodwin K. D., Thomson L.R., Duarte B., Kahlke T., Thompson A. R., Marques J. C., Cacador I. (2017). DNA Sequencing as a tool to monitor marine ecological status. *Front. Mar. Sci.* **4**:107

Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R, *et al.* (2008). Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environ Microbiol* **10**: 3349–3365.

Hahn MW. (2004). Broad diversity of viable bacteria in 'sterile' (0.2 μm) filtered water. *Res Microbiol* **155**: 688–691.

Handelsman J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68**: 669–685.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5**: R245–R249.

Hasegawa H, Naganuma K, Nakagawa Y, Matsuyama T. (2003). Membrane filter (pore size, 0.22^0.45 um; thickness, 150 um) passing-through activity of Pseudomonas aeruginosa and other bacterial species with indigenous infiltration ability. *FEMS Microbiol Lett* **223**: 41–46.

Herndl GJ, Reinthaler T. (2013). Microbial control of the dark end of the biological pump. *Nat Geosci* **6**: 718–724.

Hingamp P, Grimsley N, Acinas SG, Clerissi C, Subirana L, Poulain J, *et al.* (2013). Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J* **7**: 1678–95.

Hoeijmakers W a M, Bártfai R, Françoijs K-J, Stunnenberg HG. (2011). Linear amplification for deep sequencing. *Nat Protoc* **6**: 1026–1036.

Holligan PM. (1992). Do Marine Phytoplankton Influence Global Climate? In: *Primary Productivity and Biogeochemical Cycles in the Sea*. Springer US: Boston, MA, pp 487–501.

Horiguchi T. (2015). Diversity and Phylogeny of Marine Parasitic Dinoflagellates. In: Ohtsuka S, Suzaki T, Horiguchi T, Suzuki N, Not F (eds). *Marine Protists*. Springer Japan: Tokyo, pp 397–419.

Houghton JT. (1996). Climate change 1995: The science of climate change. Edited by J. T. Houghton, L. G. Meira Filho, B. A. Callander, N. Harris, A. Katenberg and K. Maskell. *Weather* **51**: 393–393.

Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield D a, *et al.* (2007). Microbial Population Structures in the Deep Marine Biosphere. *Science (80-)* **318**: 97–100.

Hugenholtz P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3(2)**: reviews0003.1–reviews0003.8.

Hunt DE, Lin Y, Church MJ, Karl DM, Tringe SG, Izzo LK, *et al.* (2013). Relationship between Abundance and Specific Activity of Bacterioplankton in Open Ocean Surface Waters. *Appl Environ Microbiol* **79**: 177–184.

Hunter RL, Halanych KM. (2008). Evaluating connectivity in the brooding brittle star Astrotoma agassizii across the drake passage in the Southern Ocean. *J Hered* **99**: 137–48.

Hurwitz BL, Sullivan MB. (2013). The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* **8**: e57355.

Iizuka T, Yamanaka S, Nishiyama T, Hiraishi A. (1998). Isolation and phylogenetic analysis of aerobic copiotrophic ultramicrobacteria from urban soil. *J Gen Appl Microbiol* **44**: 75–84.

Ikeda K, Doi H, Tanaka K, Kawai T, Negishi JN. (2016). Using environmental DNA to detect an endangered crayfish Cambaroides japonicus in streams. *Conserv Genet Resour* **8**: 231–234.

Ikeda Y, Siedler G, Zwierz M. (1989). On the variability of Southern Ocean front locations between southern Brazil and the Antarctic Peninsula. *J Geophys Res* **94**: 4757.

Iyer LM, Aravind L, Koonin E V. (2001). Common Origin of Four Diverse Families of Large Eukaryotic DNA Viruses. *J Virol* **75**: 11720–11734.

John SG, Mendez CB, Deng L, Poulos B, Kauffman AKM, Kern S, *et al.* (2011). A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ Microbiol Rep* **3**: 195–202.

Koonin E V, Senkevich TG, Dolja V V. (2006). The ancient Virus World and evolution of cells. *Biol Direct* **1**: 29.

Koonin E V, Yutin N. (2010). Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. *Intervirology* **53**: 284–92.

Kopf A, Bicak M, Kottmann R, Schnetzer J, Kostadinov I, Lehmann K, *et al.* (2015). The ocean sampling day consortium. *Gigascience* **4**: 27.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. (2008). A bioinformatician’s guide to metagenomics. *Microbiol Mol Biol Rev* **72**: 557–78, Table of Contents.

Laursen L. (2011). Spain’s ship comes in. *Nature* **475**: 16–17.

Lavigne R, Darius P, Summer EJ, Seto D, Mahadevan P, Nilsson AS, *et al.* (2009). Classification of Myoviridae bacteriophages using protein sequence similarity. *BMC Microbiol* **9**: 224.

Lawrence JE, Steward GF. (2010). Purification of viruses by centrifugation. *Man Aquat Viral Ecol ASLO* **2**: 166–181.

Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, *et al.* (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc Natl Acad Sci* **111**: 201320670.

Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot M, *et al.* (2015). In-depth study of Mollivirus sibericum , a new 30,000-y-old giant virus infecting Acanthamoeba. *Proc Natl Acad Sci* **112**: 201510795.

Lessard EJ, Swift E. (1986). Dinoflagellates from the North Atlantic c1assified as phototrophic or heterotrophic by epifluorescence microscopy. *J Plankton Res* **8**: 1209–1215.

Li W, Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Li WKW. (2009). From cytometry to macroecology: A quarter century quest in microbial oceanography. *Aquat Microb Ecol* **57**: 239–251.

Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, *et al.* (2015). Determinants of community structure in the global plankton interactome. *Science (80-)* **348**: 1262073–1262073.

Liu S, Vijayendran D, Bonning BC. (2011). Next generation sequencing technologies for insect virus discovery. *Viruses* **3**: 1849–69.

Loecher M, Ropkins K. (2015). RgoogleMaps and loa : Unleashing R Graphics Power on Map Tiles. *J Stat Softw* **63**: 1–18.

Logares R, Haverkamp THA, Kumar S, Lanzén A, Nederbragt AJ, Quince C, *et al.* (2012). Environmental microbiology through the lens of high-throughput DNA sequencing: Synopsis of current platforms and bioinformatics approaches. *J Microbiol Methods* **91**: 106–113.

Loman NJ, Constantinidou C, Chan JZM, Halachev M, Sergeant M, Penn CW, *et al.* (2012a). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol* **10**: 599–606.

Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, *et al.* (2012b). Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* **30**: 434–439.

Longhurst AR, Glen Harrison W. (1989). The biological pump: Profiles of plankton production and consumption in the upper ocean. *Prog Oceanogr* **22**: 47–123.

Ludwig W, Klenk H-P. Overview: A Phylogenetic Backbone and Taxonomic Framework for Procaryotic Systematics. Bergey's manual® of systematic bacteriology, 2001 - Springer

Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* **7**: e30087.

Lutjeharms JRE, de Ruijter WPM. (1996). The influence of the Agulhas Current on the adjacent coastal ocean: possible impacts of climate change. *J Mar Syst* **7**: 321–336.

Macdonald AM, Wunsch C. (1996). An estimate of global ocean circulation and heat fluxes. *Nature* **382**: 436–439.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2**: e593.

Malviya S, Scalco E, Audic S, Vincent F, Veluchamy A, Bittner L, *et al.* (2015). Insights into global diatom distribution and diversity in the world's ocean. *Proc Natl Acad Sci* **348**: in review.

Margalef R. (1969). Perspectives in ecological theory. Univ. Chicago Press, Chicago, Ill. 111 p, Limnology and Oceanography, 14, doi: 10.4319/lo.1969.14.2.0312a.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–80.

Margulis L, Sagan D. (1997). Microcosmos : four billion years of evolution from our microbial ancestors. University of California Press.

Marine R, Polson SW, Ravel J, Hatfull G, Russell D, Sullivan M, *et al.* (2011). Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl Environ Microbiol* **77**: 8071–8079.

Martínez Martínez J, Swan BK, Wilson WH. (2014). Marine viruses, a genetic reservoir revealed by targeted viromics. *ISME J* **8**: 1079–88.

Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, *et al.* (2006). Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**: 102–112.

Massana R. (2011). Eukaryotic picoplankton in surface oceans. *Annu Rev Microbiol* **65**: 91–110.

Massana R, Pedrós-Alió C. (2008). Unveiling new microbial eukaryotes in the surface ocean. *Curr Opin Microbiol* **11**: 213–218.

Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, *et al.* (2016). Linking Virus Genomes with Host Taxonomy. *Viruses* **8**: 66.

Millard AD, Zwirglmaier K, Downey MJ, Mann NH, Scanlan DJ. (2009). Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of Synechococcus host genes localized to a hyperplastic region: Implications for mechanisms of cyanophage evolution. *Environ Microbiol* **11**: 2370–2387.

Mineta K, Gojobori T. (2016). Databases of the marine metagenomics. *Gene* **576**: 724–728.

Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. (2013). Expanding the Marine Virosphere Using Metagenomics. *PLoS Genet* **9**. e-pub ahead of print, doi: 10.1371/journal.pgen.1003987.

Monier A, Claverie J-M, Ogata H. (2008). Taxonomic distribution of large DNA viruses in the sea. *Genome Biol* **9**: R106.

Moore CM, Mills MM, Arrigo KR, Berman-Frank I, Bopp L, Boyd PW, *et al.* (2013). Processes and patterns of oceanic nutrient limitation. *Nat Geosci* **6**: 701–710.

Moxley K, Schmidt S. (2012). Isolation of a phenol-utilizing marine bacterium from Durban Harbour (South Africa) and its preliminary characterization as Marinobacter sp. KM2. *Water Sci Technol* **65**: 932–9.

Munn CB. (2006). Viruses as pathogens of marine organisms—from bacteria to whales. *J Mar Biol Assoc UK* **86**: 453.

Nagasaki K. (2008). Dinoflagellates, diatoms, and their viruses. *J Microbiol* **46**: 235–243.

Nagasaki K, Tomaru Y, Shirai Y, Takao Y, Mizumoto H. (2006). Dinoflagellate-infecting viruses. *J Mar Biol Assoc UK* **86**: 469.

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, *et al.* (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* **39**: e90–e90.

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA. (1986). Microbial Ecology and Evolution: A Ribosomal RNA Approach. *Annu Rev Microbiol* **40**: 337–365.

Ondov BD, Bergman NH, Phillippy AM. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**: 385.

Pace NR. (1997). A molecular view of microbial diversity and the biosphere. *Science* **276**: 734–740.

Pace NR, Stahl DA, Lane DJ, Olsen GJ. (1986). The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. In: Springer US, pp 1–55.

Palumbi SR. (1994). Genetic Divergence, Reproductive Isolation, and Marine Speciation. *Annu Rev Ecol Syst* **25**: 547–572.

Palumbi SR. (1992). Marine speciation on a small planet. *Trends Ecol Evol* **7**: 114–118.

Pasteur L. (1885). Observations relatives à la note précédente... - Google Scholar. *Compte Rendus Ge Acad Sci* **100**: 68.

Pedrós-Alió C. (2006). Marine microbial diversity: can it be determined? *Trends Microbiol* **14**: 257–263.

Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, *et al.* (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**: 171–182.

Perez Sepulveda B, Redgwell T, Rihtman B, Pitt F, Scanlan DJ, Millard A. (2016). Marine phage genomics: the tip of the iceberg Hantke K (ed). *FEMS Microbiol Lett* **363**: fnw158.

Pfaff M, Flaviani F, Du Plessis G, Rybycki E, Schroeder D. (2014). Research and Technology Research and Technology. In: Funke, N., Claassen, M., Meissner, R. and Nortje K (ed). *Reflections on the State of Research and Development in the Marine and Maritime Sectors in South Africa*. The Council for Scientific and Industrial Research Pretoria, South Africa, pp 100–121.

Pfennig N. (1967). Photosynthetic Bacteria. *Annu Rev Microbiol* **21**: 285–324.

Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, *et al.* (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**: 281–6.

Polz MF, Cavanaugh CM. (1998). Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* **64**: 3724–30.

Popova EE, Ryabchenko VA, Fasham MJR. (2000). Biological pump and vertical mixing in the southern ocean: Their impact on atmospheric CO 2. *Global Biogeochem Cycles* **14**: 477–498.

Proctor LM, Fuhrman JA. (1990). Viral mortality of marine bacteria and cyanobacteria. *Nature* **343**: 60–62.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, *et al.* (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Rahmstorf S. (2002). Ocean circulation and climate during the past 120,000 years. *Nature* **419**: 207–14.

Rappé MS, Giovannoni SJ. (2003). The uncultured microbial majority. *Annu Rev Microbiol* **57**: 369–94.

Reeder J, Knight R. (2009). The 'rare biosphere': a reality check. *Nat Methods* **6**: 636–637.

Reinthaler T, Van Aken H, Veth C, Arístegui J, Robinson C, Williams PJ. leB ., *et al.* (2009). Prokaryotic respiration and production in the meso- and bathypelagic realm of the eastern and western North Atlantic basin. *Limnol Oceanogr* **51**: 1262–1273.

Rio M-H, Mulet S, Picot N. (2013). New global mean dynamic topography from a goce geoid model, altimeter measurements and oceanographic in-situ data. In: *ESA Living Planet Symposium, Proceedings of the conference held on 9-13 September 2013 at Edinburgh in United Kingdom. ESA SP-722. 2-13, p.27*. pp 2–13.

Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD, *et al.* (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**: 283–290.

Rohwer F, Thurber RV. (2009). Viruses manipulate the marine environment. *Nature* **459**: 207–12.

Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D, *et al.* (2011). Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27**: 3074–5.

Roux S, Hallam SJ, Woyke T, Sullivan MB. (2015). Viral dark matter and virus – host interactions resolved from publicly available microbial genomes. *Elife* **4**: 1–20.

Roux S, Krupovic M, Debroas D, Forterre P, Enault F. (2013). Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol* **3**: 130160.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.

Salcher MM, Pernthaler J, Frater N, Posch T. (2011). Vertical and longitudinal distribution patterns of different bacterioplankton populations in a canyon-shaped, deep prealpine lake. *Limnol Oceanogr* **56**: 2027–2039.

Sandaa R. (2008). Burden or benefit? Virus-host interactions in the marine environment. *Res Microbiol* **159**: 374–81.

Sanders RW, Berninger UG, Lim EL, Kemp PF, Caron DA. (2000). Heterotrophic and mixotrophic nanoplankton predation on picoplankton in the Sargasso Sea and on Georges Bank. *Mar Ecol Prog Ser* **192**: 103–118.

Sano E, Carlson S, Wegley L, Rohwer F. (2004). Movement of viruses between biomes. *Appl Environ Microbiol* **70**: 5842–6.

Schmidt P, Bálint M, Greshake B, Bandow C, Römbke J, Schmitt I. (2013). Illumina metabarcoding of a soil fungal community. *Soil Biol Biochem* **65**: 128–132.

Scholz MB, Lo C, Chain PSG. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr Opin Biotechnol* **23**: 9–15.

Seymour JR, Seuront L, Doubell M, Waters RL, Mitchell JG. (2006). Microscale patchiness of virioplankton. *J Mar Biol Assoc UK* **86**: 551.

Shaw PW, Arkhipkin AI, Al-khairullaI H. (2004). Genetic structuring of Patagonian toothfish populations in the Southwest Atlantic Ocean: the effect of the Antarctic Polar Front and deep-water troughs as barriers to genetic exchange. *Mol Ecol* **13**: 3293–3303.

Sherr E, Sherr B. (2000). Marine microbes: an overview. *Microb Ecol Ocean* 13–46.

Sieburth JM, Smetacek V, Lenz J. (1978). Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions 1. *Limnol Oceanogr* **23**: 1256–1263.

Simpson AGB, Roger AJ. (2004). The real ‘kingdoms’ of eukaryotes. *Curr Biol* **14**: R693–R696.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* **15**: 121–32.

Smayda TJ. (1997). Harmful algal blooms: Their ecophysiology and general relevance to phytoplankton blooms in the sea. *Limnol Oceanogr* **42**: 1137–1153.

Smit P, Heniger J. (1975). Antoni van Leeuwenhoek (1632–1723) and the discovery of bacteria. *Antonie Van Leeuwenhoek* **41**: 217–228.

Sobecky PA, Hazen TH. (2009). Horizontal gene transfer and mobile genetic elements in marine systems. *Methods Mol Biol* **532**: 435–453.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, *et al.* (2006). Microbial diversity in the deep sea and the underexplored ‘rare biosphere’. *Proc Natl Acad Sci* **103**: 12115–12120.

Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, *et al.* (2007). IPCC, 2007: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Solonenko SA, Ignacio-Espinoza JC, Alberti A, Cruaud C, Hallam S, Konstantinidis K, *et al.* (2013). Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics* **14**: 320.

Staley JT. (1997). Biodiversity: are microbial species threatened? *Curr Opin Biotechnol* **8**: 340–345.

Staley JT, Konopka A. (1985). Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. *Annu Rev Microbiol* **39**: 321–346.

Steward GF, Culley AI, Mueller J a, Wood-Charlson EM, Belcaid M, Poisson G. (2013). Are we missing half of the viruses in the ocean? *ISME J* **7**: 672–9.

Steward GF, Montiel JL, Azam F. (2000). Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol Oceanogr* **45**: 1697–1706.

Stoeck T, Bass D, Nebel M, Christen R, Jones M d. M, Breiner H-W, *et al.* (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* **19**: 21–31.

Sul WJ, Oliver TA, Ducklow HW, Amaral-Zettler LA, Sogin ML. (2013). Marine bacteria exhibit a bipolar distribution. *Proc Natl Acad Sci U S A* **110**: 2342–2347.

Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. (2005). Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* **3**: e144.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, *et al.* (2015). Structure and function of the global ocean microbiome. *Science (80-)* **348**: 1261359–1261359.

Suttle CA. (2007). Marine viruses — major players in the global ecosystem. *Nat Rev Microbiol* **5**: 801–812.

Suttle CA. (1994). The significance of viruses to mortality in aquatic microbial communities. *Microb Ecol* **28**: 237–243.

Suttle CA. (2005). Viruses in the sea. *Nature* **437**: 356–61.

Suttle CA, Chan AM, Cottrell MT. (1990). Infection of phytoplankton by viruses and reduction of primary productivity. *Nature* **347**: 467–469.

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH. (2012). Environmental DNA. *Mol Ecol* **21**: 1789–1793.

Tabor PS, Ohwada K, Colwell RR. (1981). Filterable marine bacteria found in the deep sea: Distribution, taxonomy, and response to starvation. *Microb Ecol* **7**: 67–83.

Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. (2014). RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res* **42**: D553–D559.

Thingstad TF. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol Oceanogr* **45**: 1320–1328.

Thomas CD, Cameron A, Green RE, Bakkenes M, Beaumont LJ, Collingham YC, *et al.* (2004). Extinction risk from climate change. *Nature* **427**: 145–8.

Thomas T, Gilbert J, Meyer F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp* **2**: 3.

Thornhill DJ, Mahon AR, Norenburg JL, Halanych KM. (2008). Open-ocean barriers to dispersal: A test case with the Antarctic Polar Front and the ribbon worm Parborlasia corrugatus (Nemertea: Lineidae). *Mol Ecol* **17**: 5104–5117.

Tilman D. (1977). Resource Competition between Plankton Algae: An Experimental and Theoretical Approach. *Ecology* **58**: 338–348.

Tomaru Y, Takao Y, Suzuki H, Nagumo T, Nagasaki K. (2009). Isolation and characterization of a single-stranded RNA virus infecting the bloom-forming diatom Chaetoceros socialis. *Appl Environ Microbiol* **75**: 2375–2381.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, *et al.* (2009). A core gut microbiome in obese and lean twins. *Nature* **457**: 480–484.

Valentini A, Pompanon F, Taberlet P. (2009). DNA barcoding for ecologists. *Trends Ecol Evol* **24**: 110–117.

Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, *et al.* (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Mol Ecol* **25**: 929–942.

de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, *et al.* (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science (80-)* **348**: 1261605–1261605.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Weinbauer MG. (2004). Ecology of prokaryotic viruses. *FEMS Microbiol Rev* **28**: 127–81.

Weynberg KD, Allen MJ, Ashelford K, Scanlan DJ, Wilson WH. (2009). From small hosts come big viruses: The complete genome of a second Ostreococcus tauri virus, OtV-1. *Environ Microbiol* **11**:

2821–2839.

Whitman WB, Coleman DC, Wiebe WJ. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95**: 6578–6583.

Wigington CH, Sonderegger D, Brussaard CPD, Buchan A, Finke JF, Fuhrman JA, *et al.* (2016). Re-examination of the relationship between marine virus and microbial cell abundances. *Nat Microbiol* **1**: 15024.

Wilhelm SW, Matteson AR. (2008). Freshwater and marine virioplankton: A brief overview of commonalities and differences. *Freshw Biol* **53**: 1076–1089.

Wilhelm SW, Suttle CA. (1999). Viruses and Nutrient Cycles in the Sea. *Bioscience* **49**: 781.

Williams KP, Gillespie JJ, Sobral BWS, Nordberg EK, Snyder EE, Shallom JM, *et al.* (2010). Phylogeny of gammaproteobacteria. *J Bacteriol* **192**: 2305–2314.

Williamson SJ, Allen LZ, Lorenzi H a, Fadrosh DW, Brami D, Thiagarajan M, *et al.* (2012). Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS One* **7**: e42047.

Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI, *et al.* (2008). The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One* **3**: e1456.

Wilson WH, Allen MJ. (2009). Giant Viruses and their Genomes. *Viral Genomes Divers Prop Parameters* 145–157.

Wilson WH, Schroeder DC, Allen MJ, Holden MTG, Parkhill J, Barrell BG, *et al.* (2005). Complete genome sequence and lytic phase transcription profile of a Coccolithovirus. *Science* **309**: 1090–2.

Woese CR, Fox GE. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74**: 5088–5090.

Wolf YI, Rogozin IB, Grishin N V, Tatusov RL, Koonin E V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol* **1**: 8.

Wommack KE, Colwell RR. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* **64**: 69–114.

Wommack KE, Sime-ngando T, Winget DM, Jamindar S, Helton RR. (2010). Filtration-based methods for the collection of viral concentrates from large water samples. *Man Aquat Viral Ecol* **12**: 110–117.

Yakimov MM, Golyshin PN, Lang S, Moore ERB, Abraham W-R, Lunsdorf H, *et al.* (1998). Alcanivorax borkumensis gen. nov., sp. nov., a new, hydrocarbon-degrading and surfactant-producing marine bacterium. *Int J Syst Bacteriol* **48**: 339–348.

Yamada T, Onimatsu H, Van Etten JL. (2006). Chlorella Viruses. In: Vol. 65. *Advances in Virus Research*. pp 293–336.

Yanai-Balser GM, Duncan GA, Eudy JD, Wang D, Li X, Agarkova I V, *et al.* (2010). Microarray analysis of Paramecium bursaria chlorella virus 1 transcription. *J Virol* **84**: 532–42.

Yolken RH, Jones-Brando L, Dunigan DD, Kannan G, Dickerson F, Severance E, *et al.* (2014). Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. *Proc Natl Acad Sci U S A* **111**: 16106–11.

Yutin N, Koonin E V. (2013). Pandoraviruses are highly derived phycodnaviruses. *Biol Direct* **8**: 25.

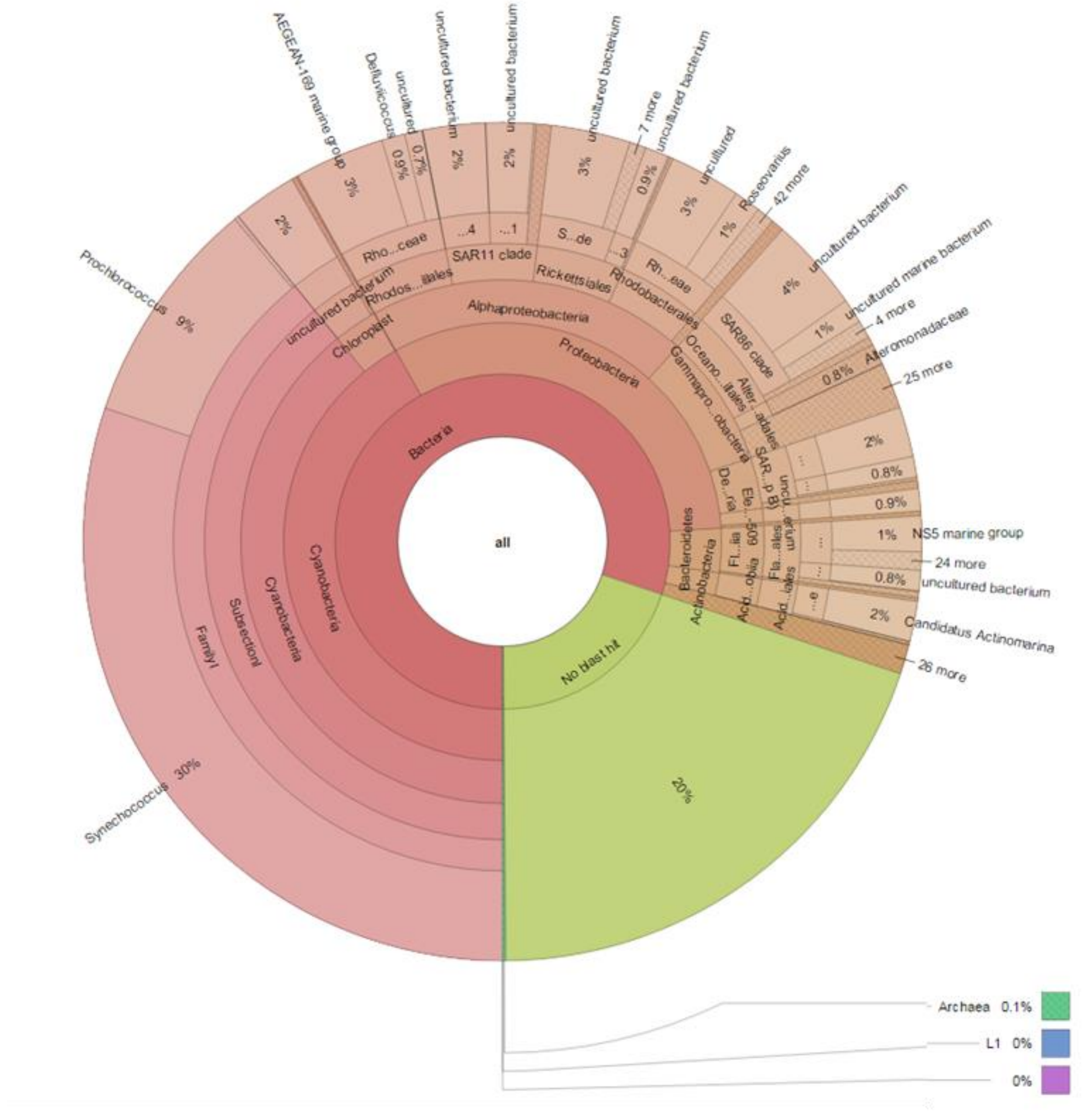
Zheng X, Dai X, Huang L. (2016). Spatial Variations of Prokaryotic Communities in Surface Water from India Ocean to Chinese Marginal Seas and their Underlining Environmental Determinants. *Front Mar Sci* **3**: 1–10.

Zhu F, Massana R, Not F, Marie D, Vaulot D. (2005). Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* **52**: 79–92.

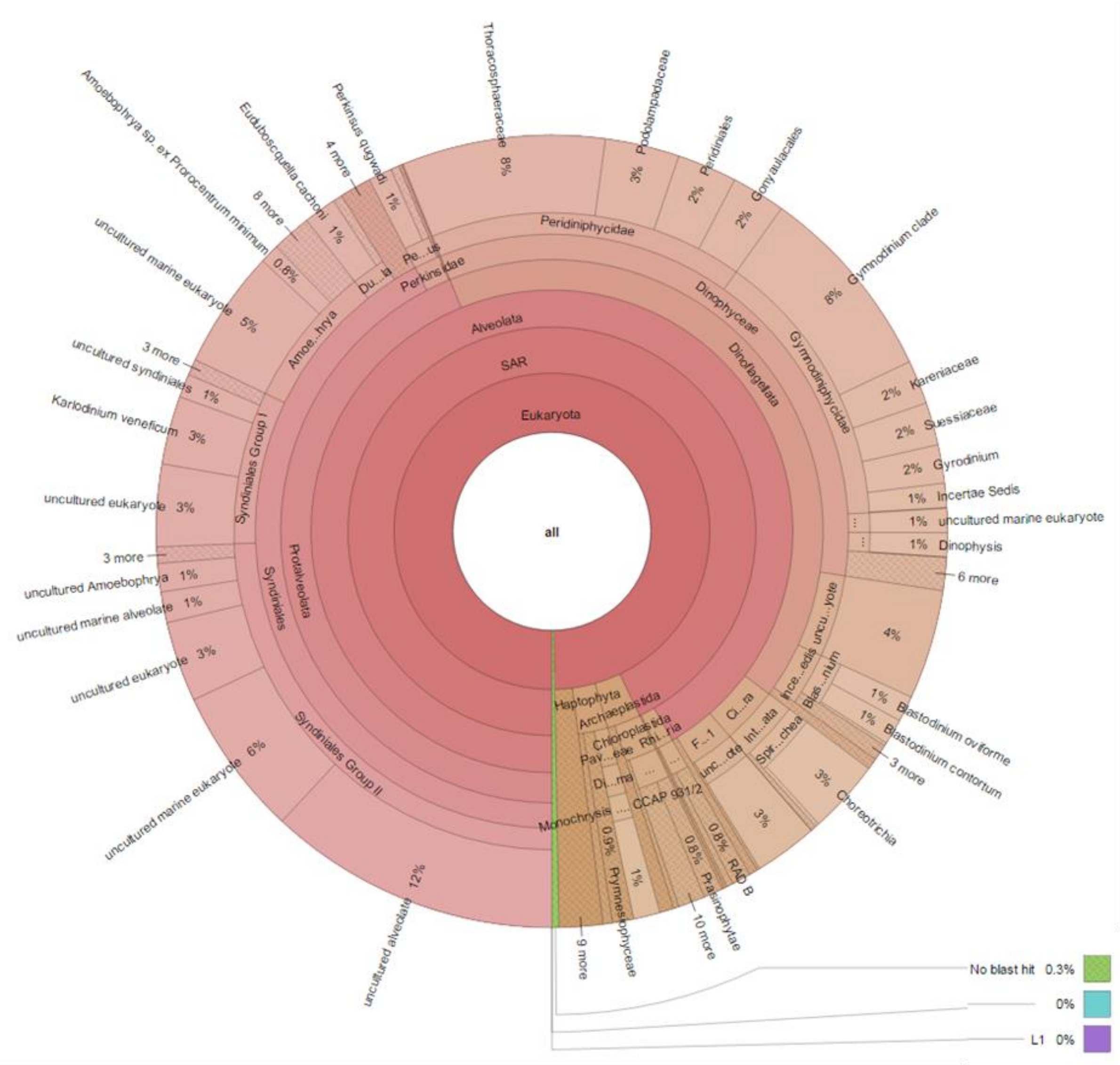
Zinger L, Amaral-Zettler LA, Fuhrman JA, Horner-Devine MC, Huse SM, Welch DBM, *et al.* (2011). Global Patterns of Bacterial Beta-Diversity in Seafloor and Seawater Ecosystems Gilbert JA (ed). *PLoS One* **6**: e24570.

Zinger L, Gobet A, Pommier T. (2012). Two decades of describing the unseen majority of aquatic microbial diversity. *Mol Ecol* **21**: 1878–1896.

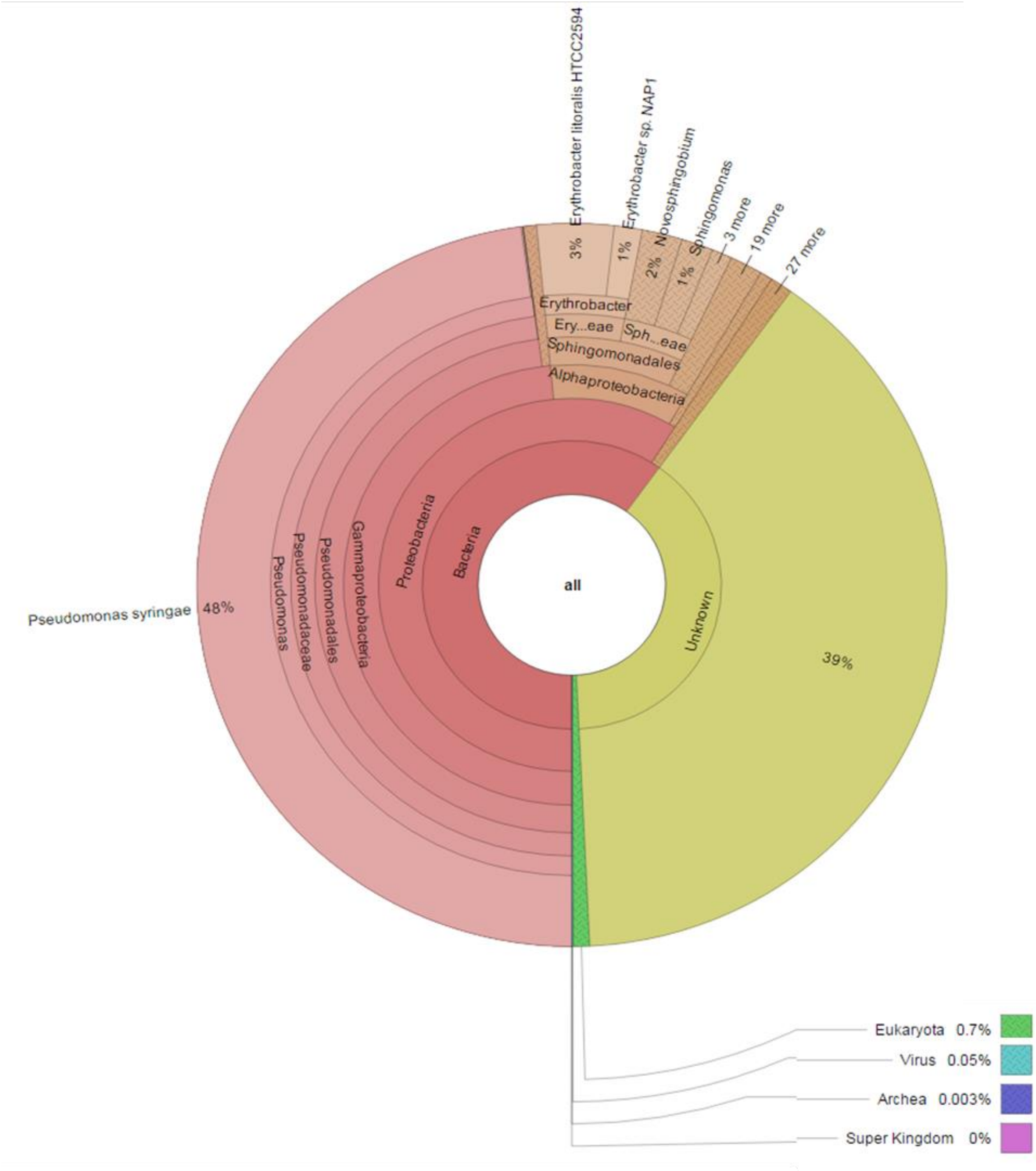
Supplementary Material: Figures



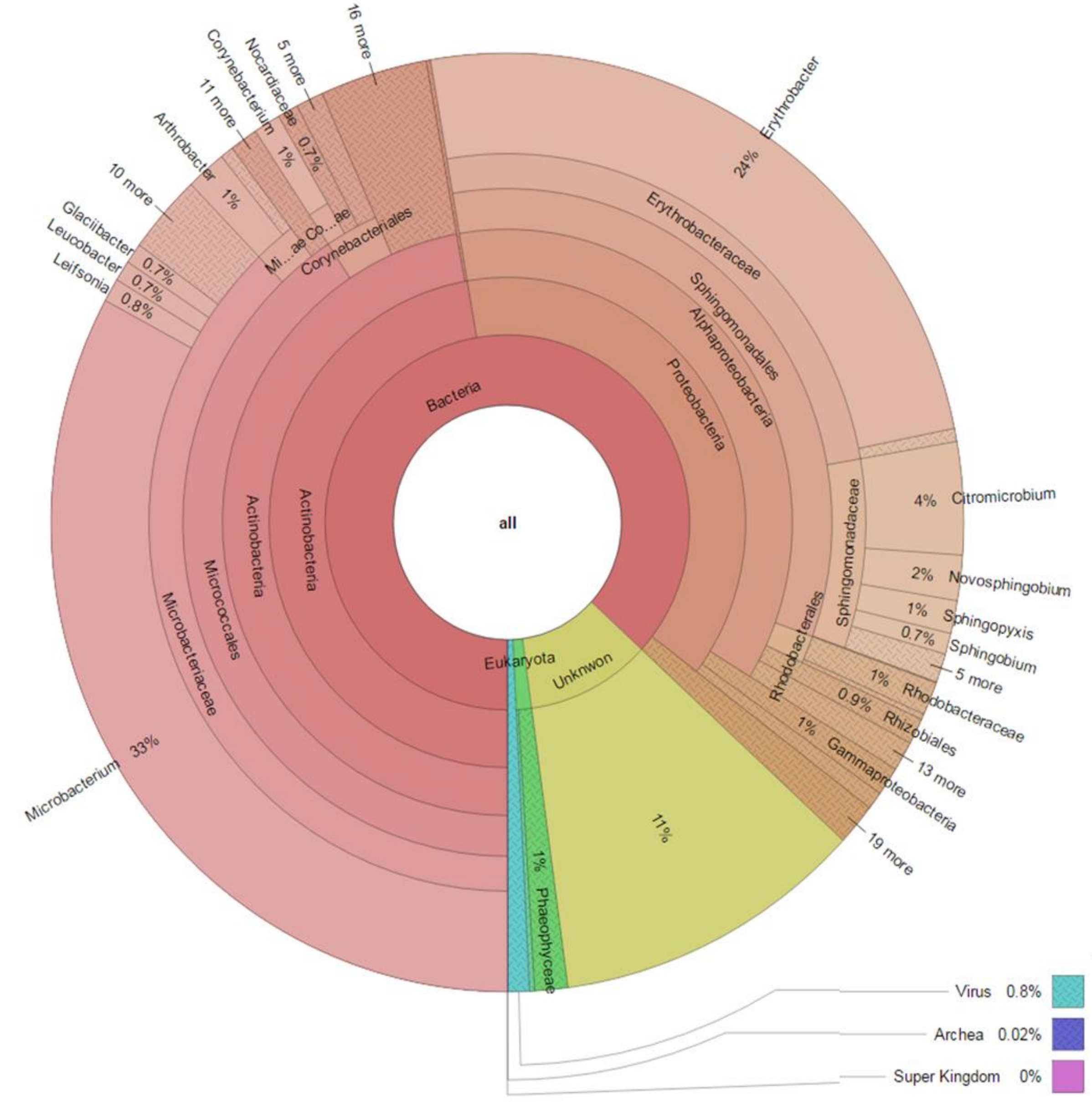
Supplementary Figure 1: Krona visualisation of prokaryotic diversity (>0.45µm fraction) based on the average of the three replicates after singleton removal (T1) in station S1.



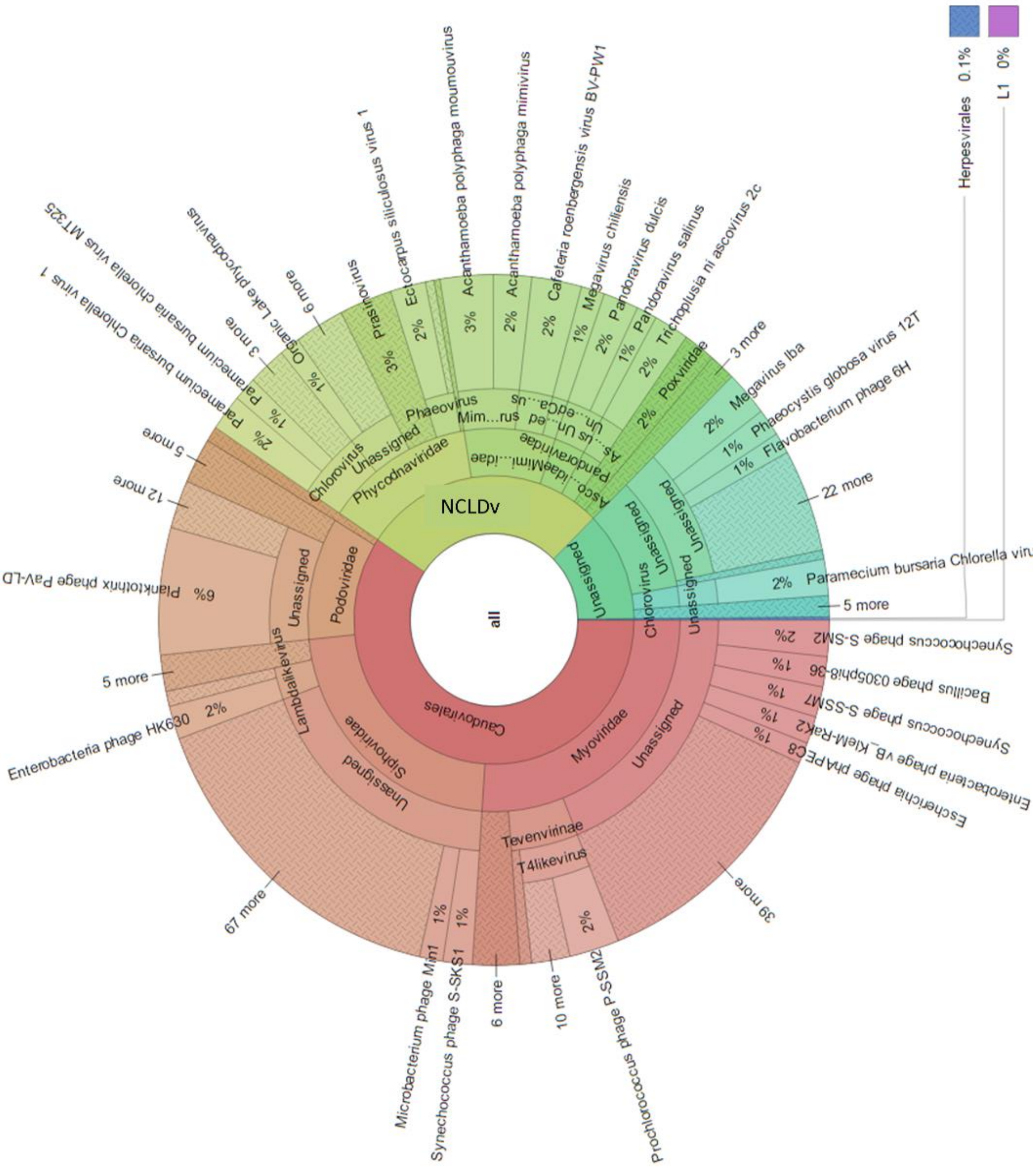
Supplementary Figure 2: Krona visualisation of the eukaryotic diversity (>0.45µm fraction) based on the average of the three replicates after singletons removal (T1) in station S1.



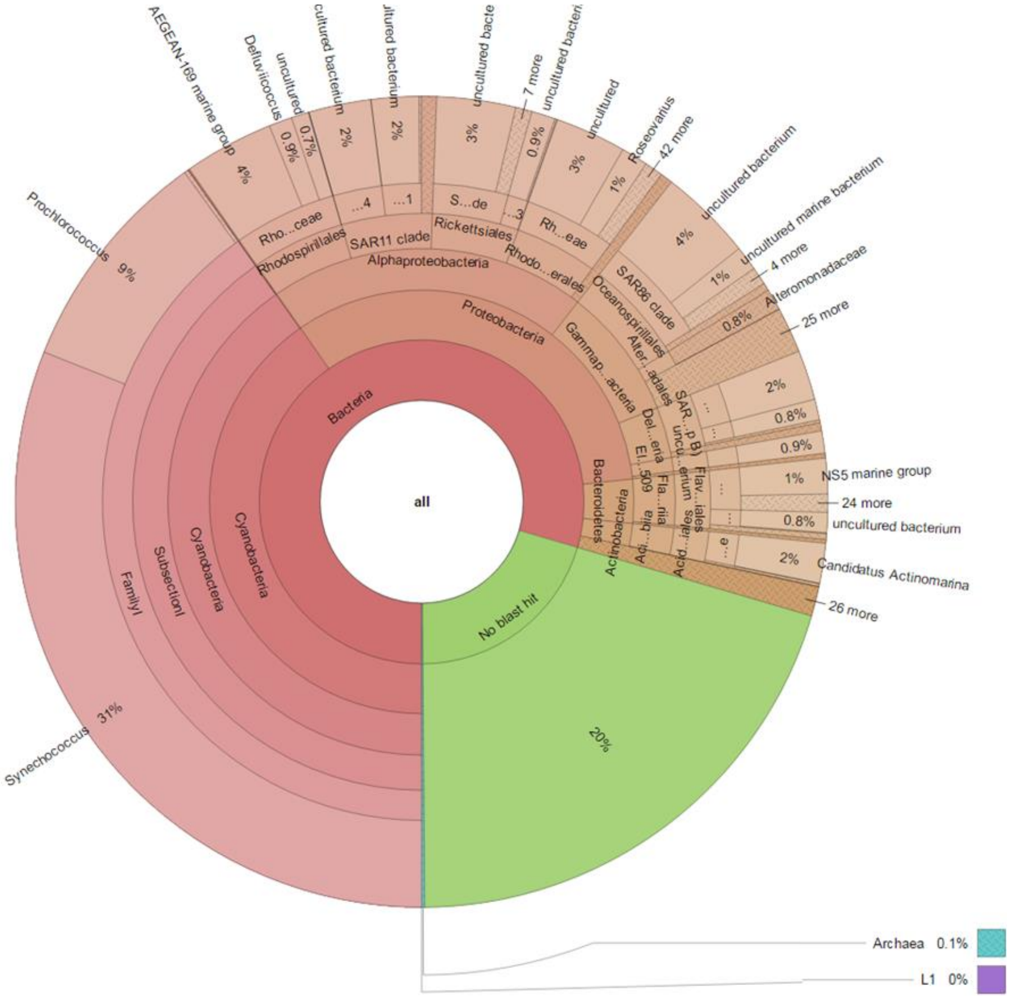
Supplementary Figure 3: krona visualisation of diversity of the <0.45µm fraction (permeate) based on Refseq annotation on R1 sub-sampled in station S1



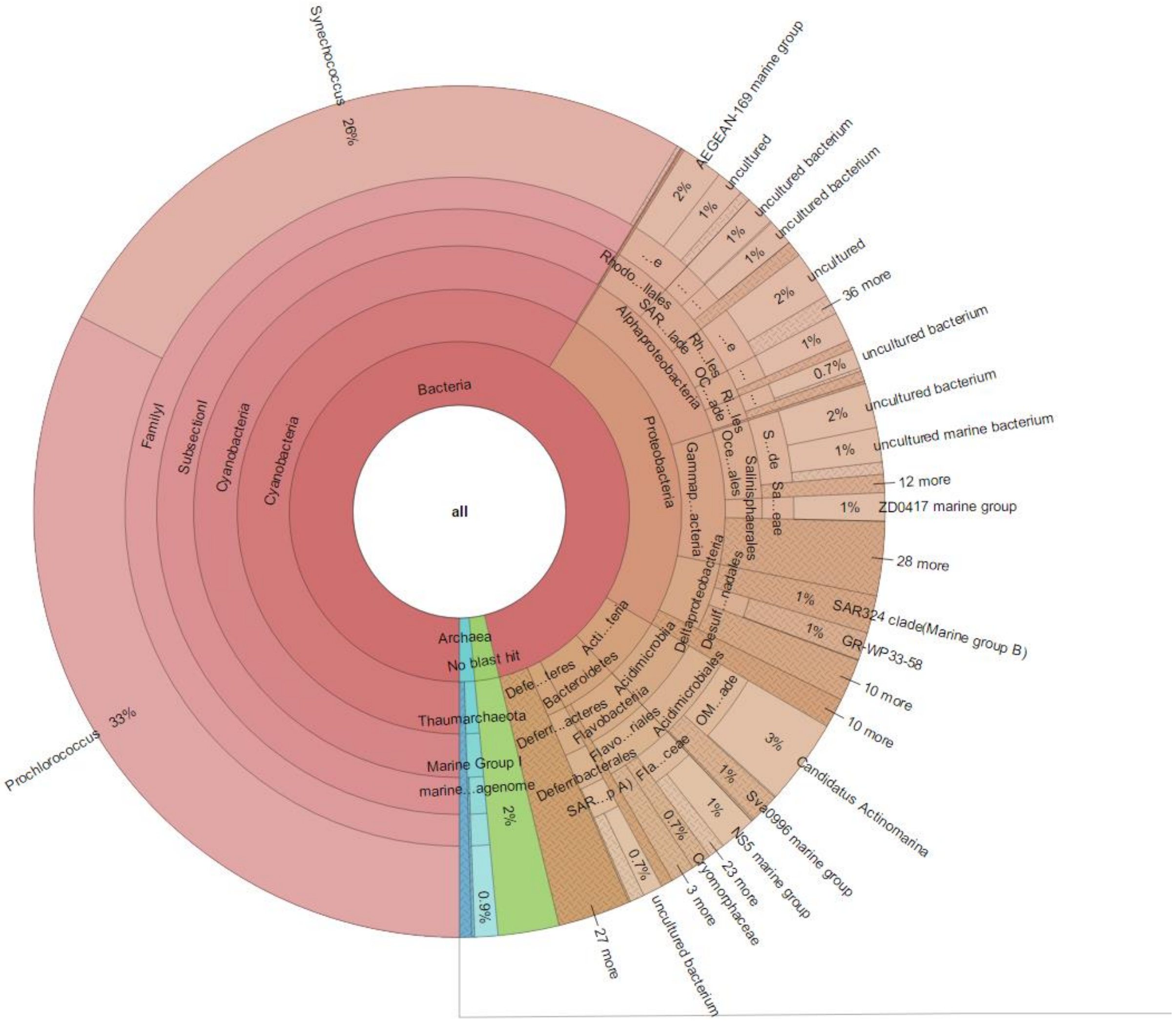
Supplementary Figure 4: krona visualisation of diversity of the <0.45µm fraction (permeate) based on Refseq annotation on the contigs in station S1



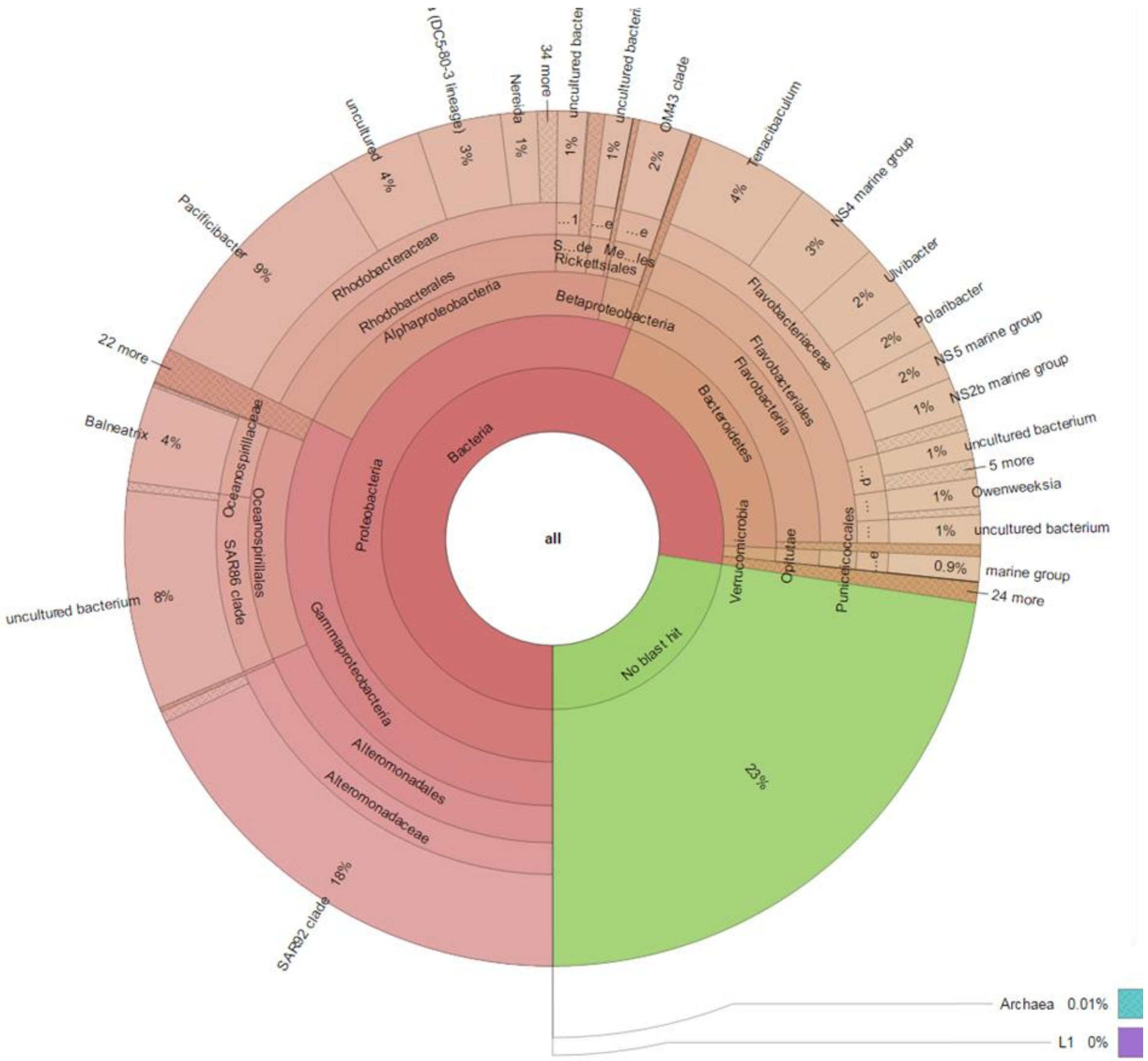
Supplementary Figure 5: krona visualisation of diversity of the <0.45µm fraction (permeate) based on Virus db annotation on the contigs in station S1



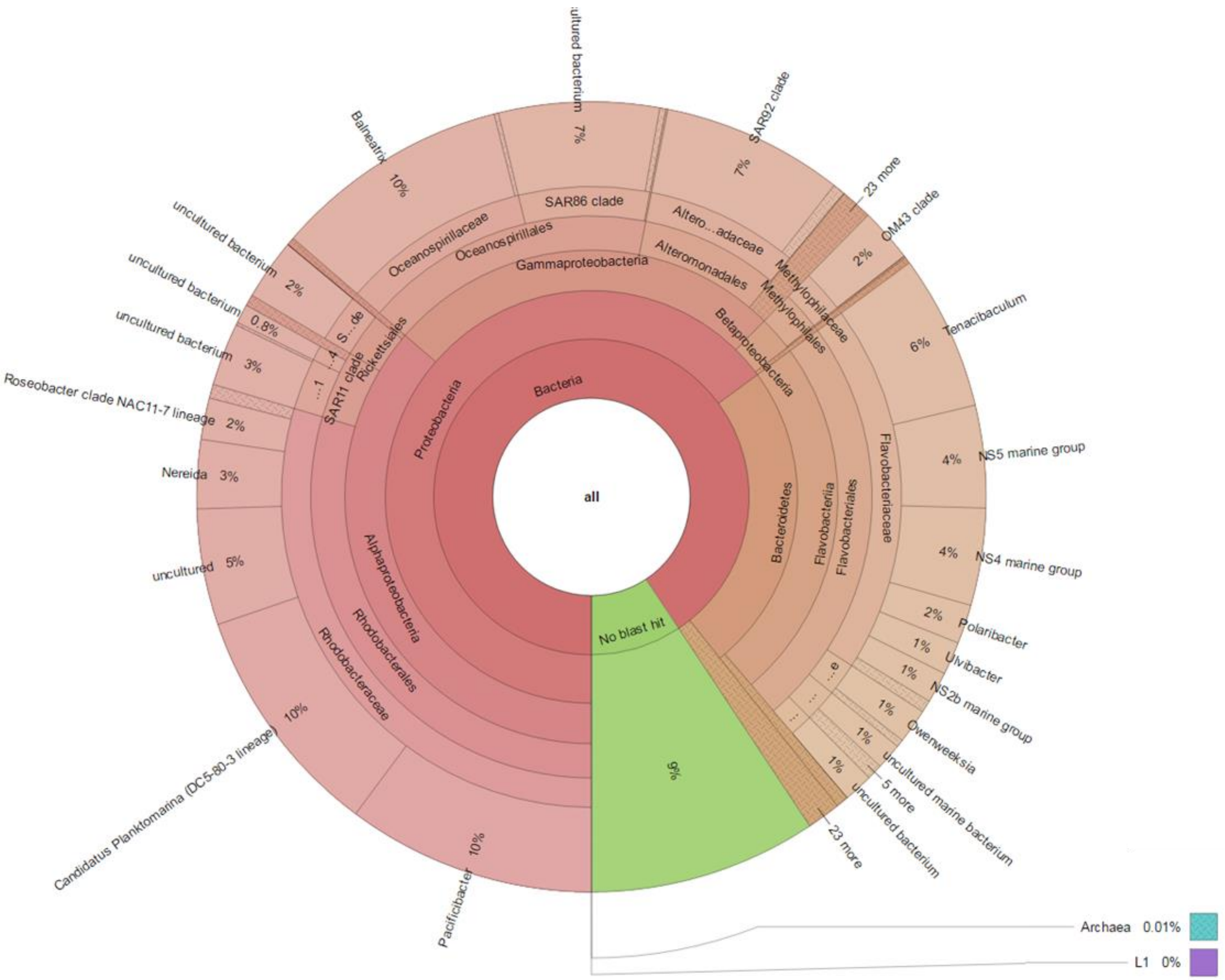
Supplementary Figure 6: Krona visualisation of the Prokaryotes in S1, showing average of the three replicates after removal of singletons, chloroplast and mitochondria.



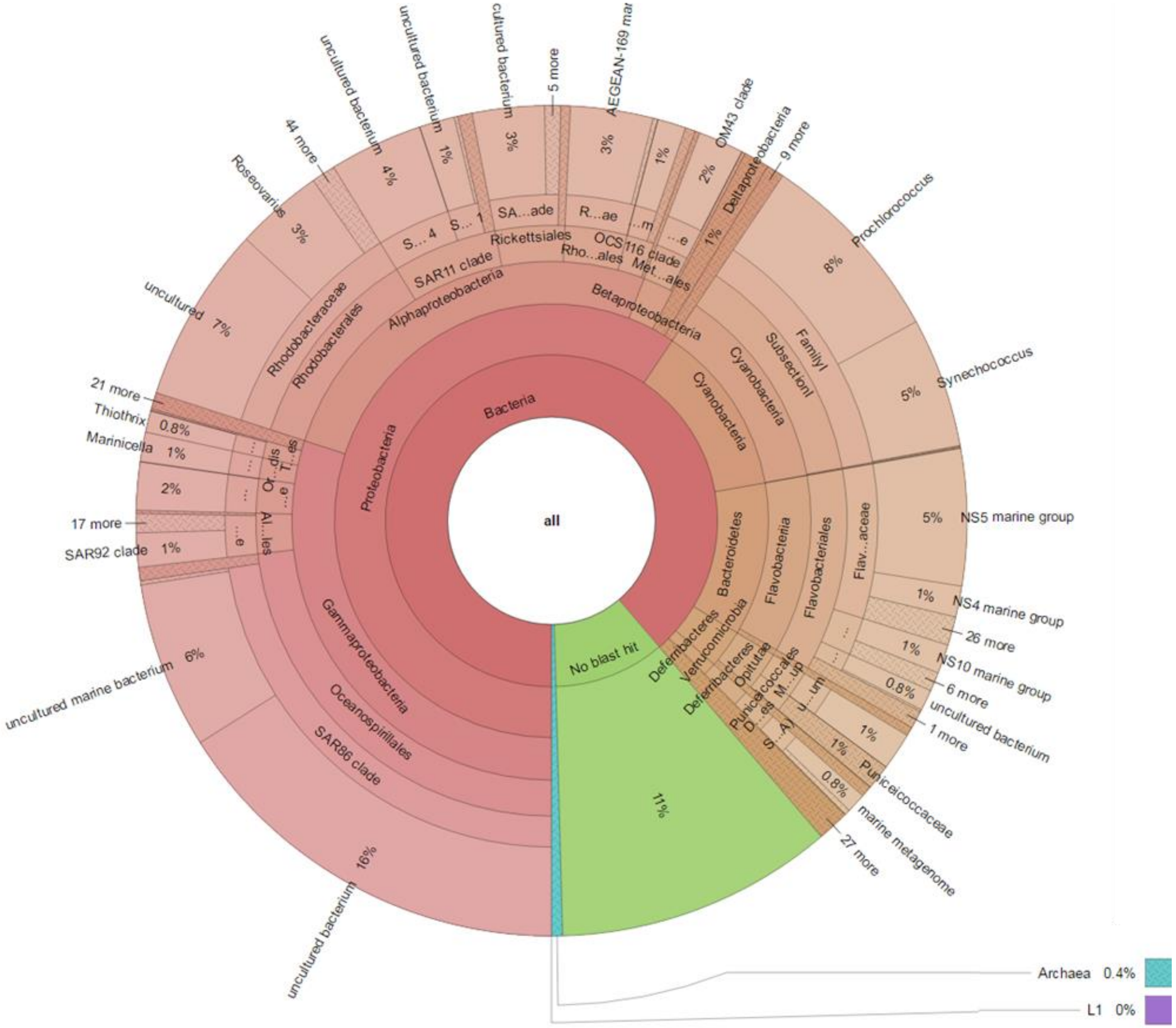
Supplementary Figure 7: Krona visualisation of the Prokaryotes in S2, showing average of the three replicates after removal of singletons, chloroplast and mitochondria.



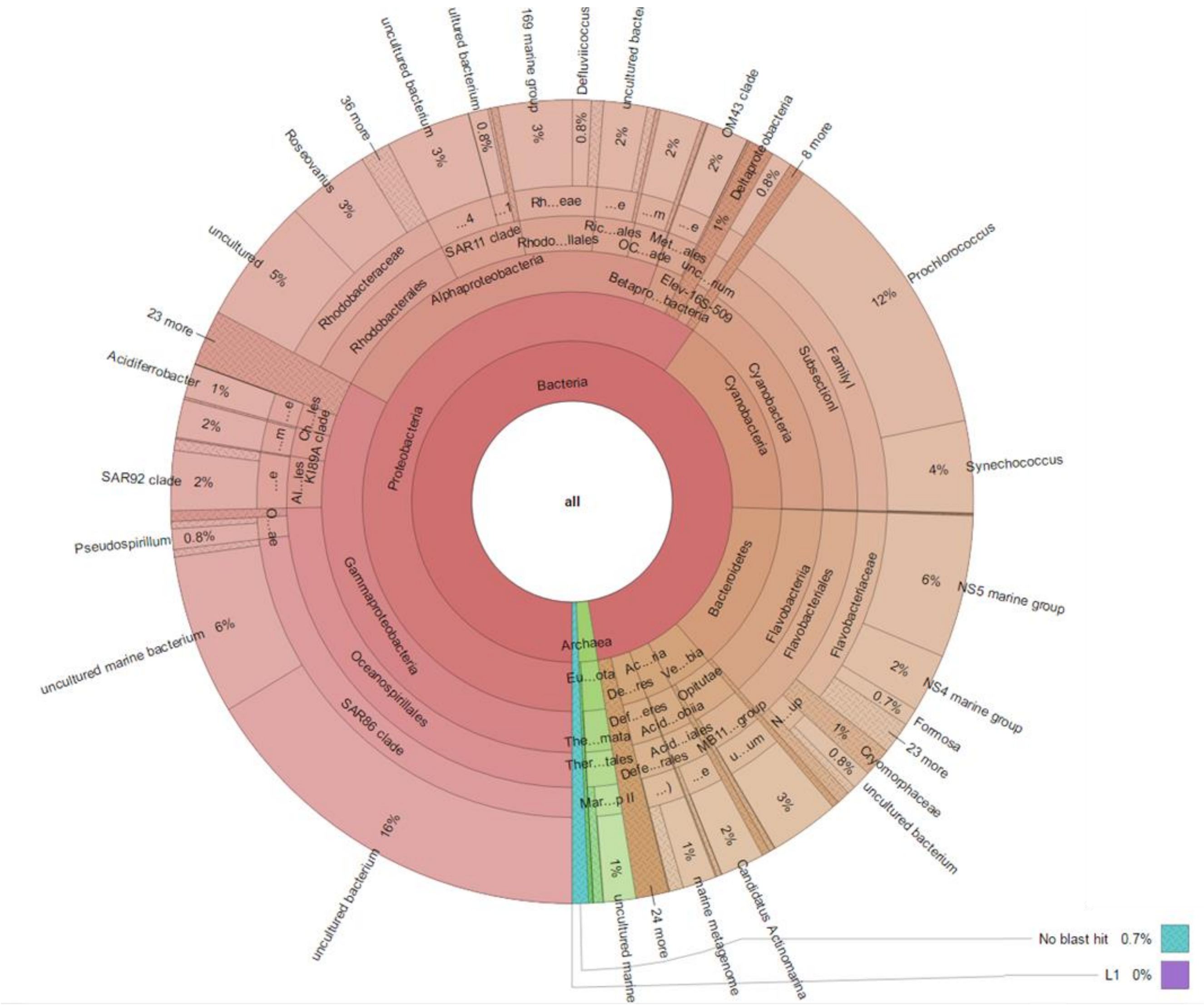
Supplementary Figure 8: Krona visualisation of the Prokaryotes in S3, showing average of the three replicates after removal of singletons, chloroplast and mitochondria.



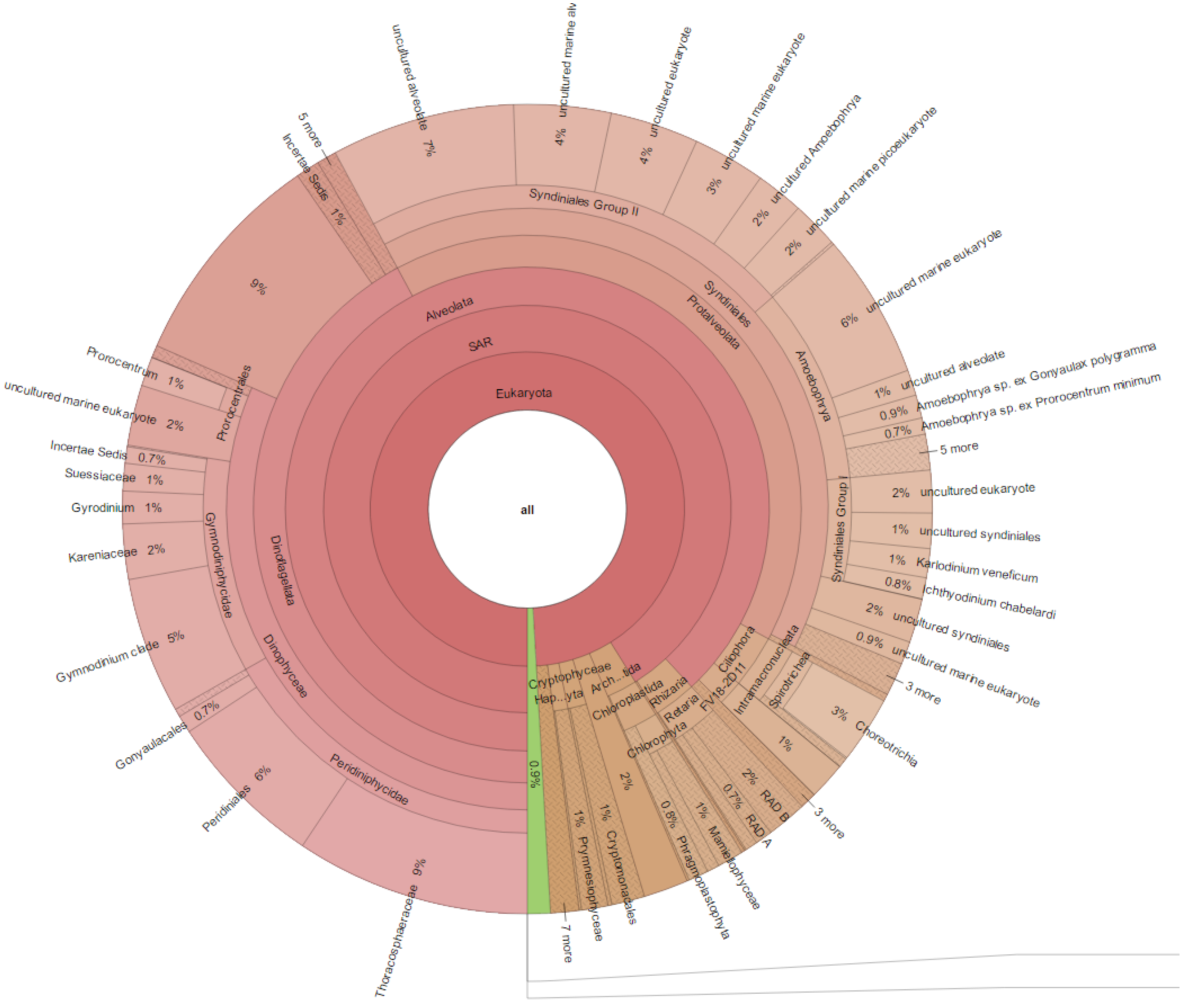
Supplementary Figure 9: Krona visualisation of the Prokaryotes in S4, showing average of the three replicates after removal of singletons, chloroplast and mitochondria.



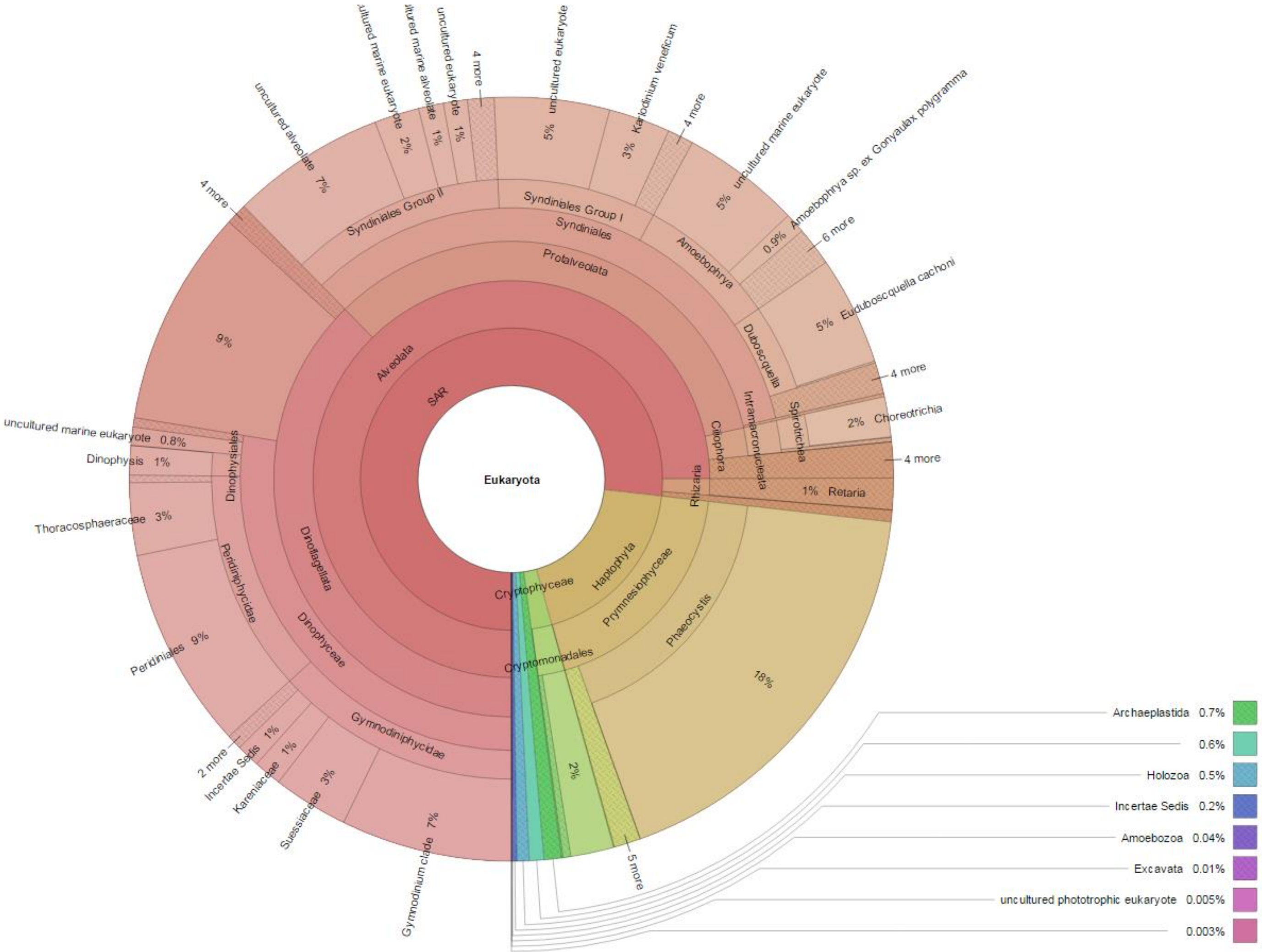
Supplementary Figure 10: Krona visualisation of the Prokaryotes in S5, showing average of the three replicates after removal of singletons, chloroplast and mitochondria.



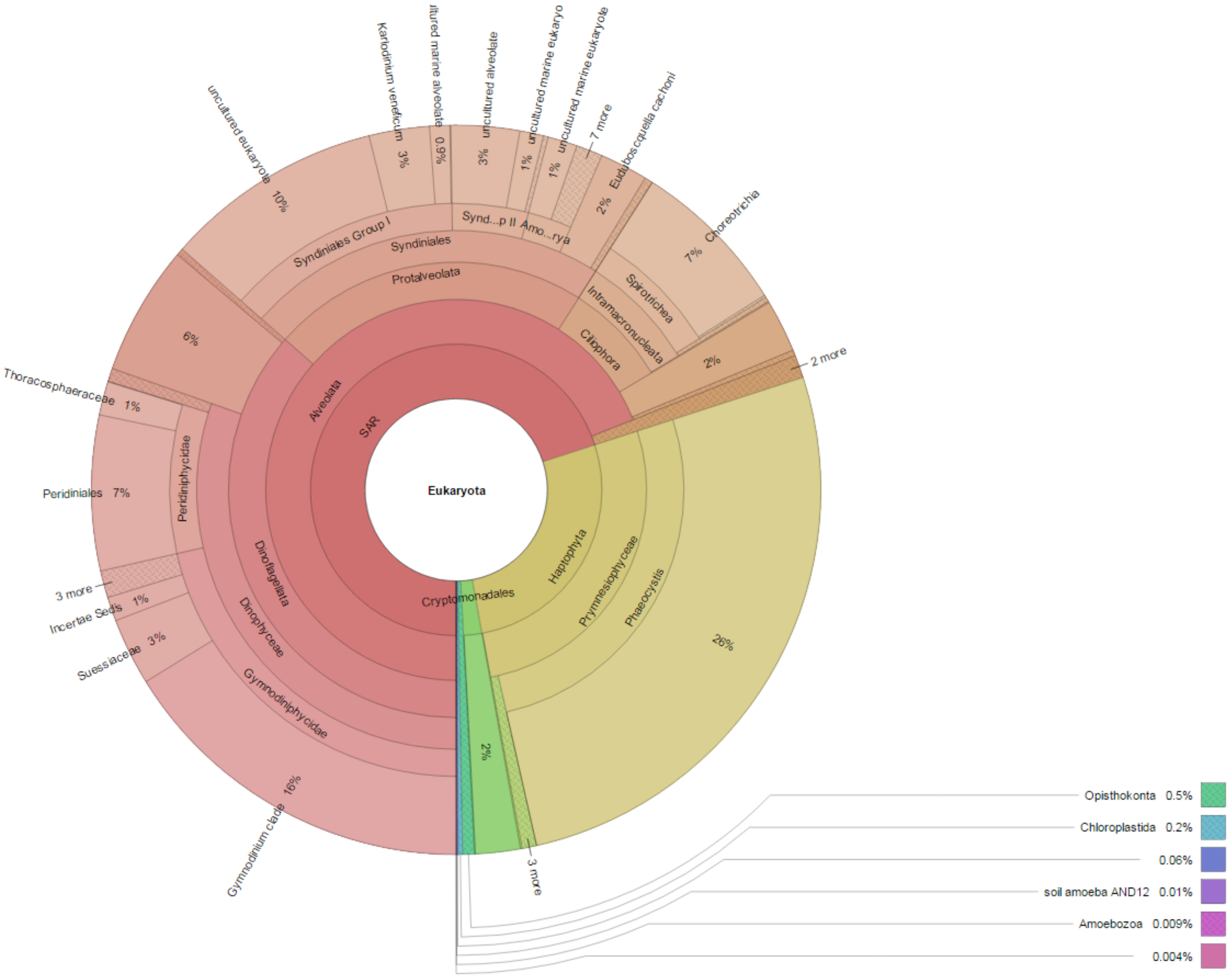
Supplementary Figure 11: Krona visualisation of the Prokaryotes in S6, showing average of the three replicates after removal of singletons, chloroplast and mitochondria.



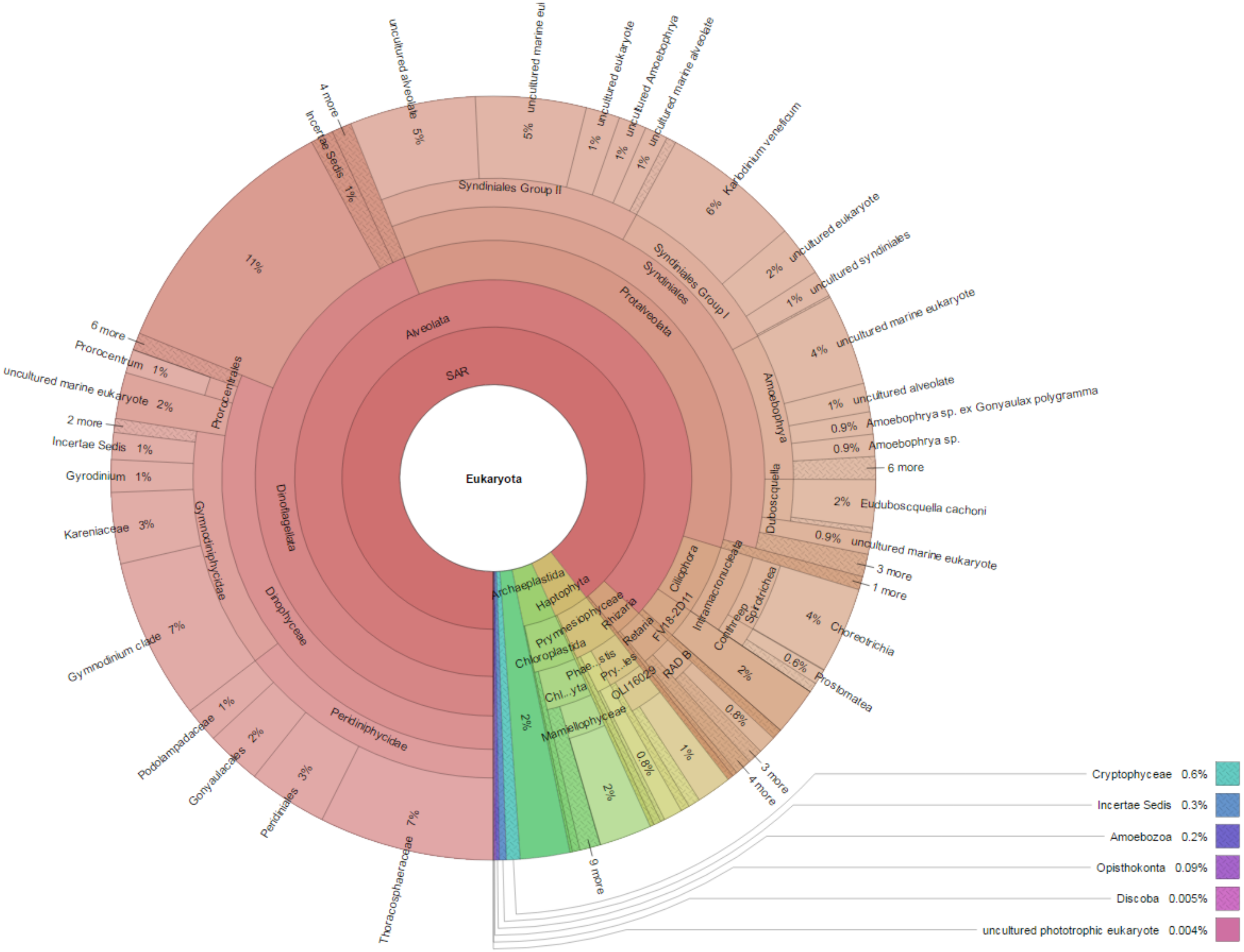
Supplementary Figure 12: Krona visualisation of the Eukaryotes in S2, showing average of the three replicates after removal of singletons.



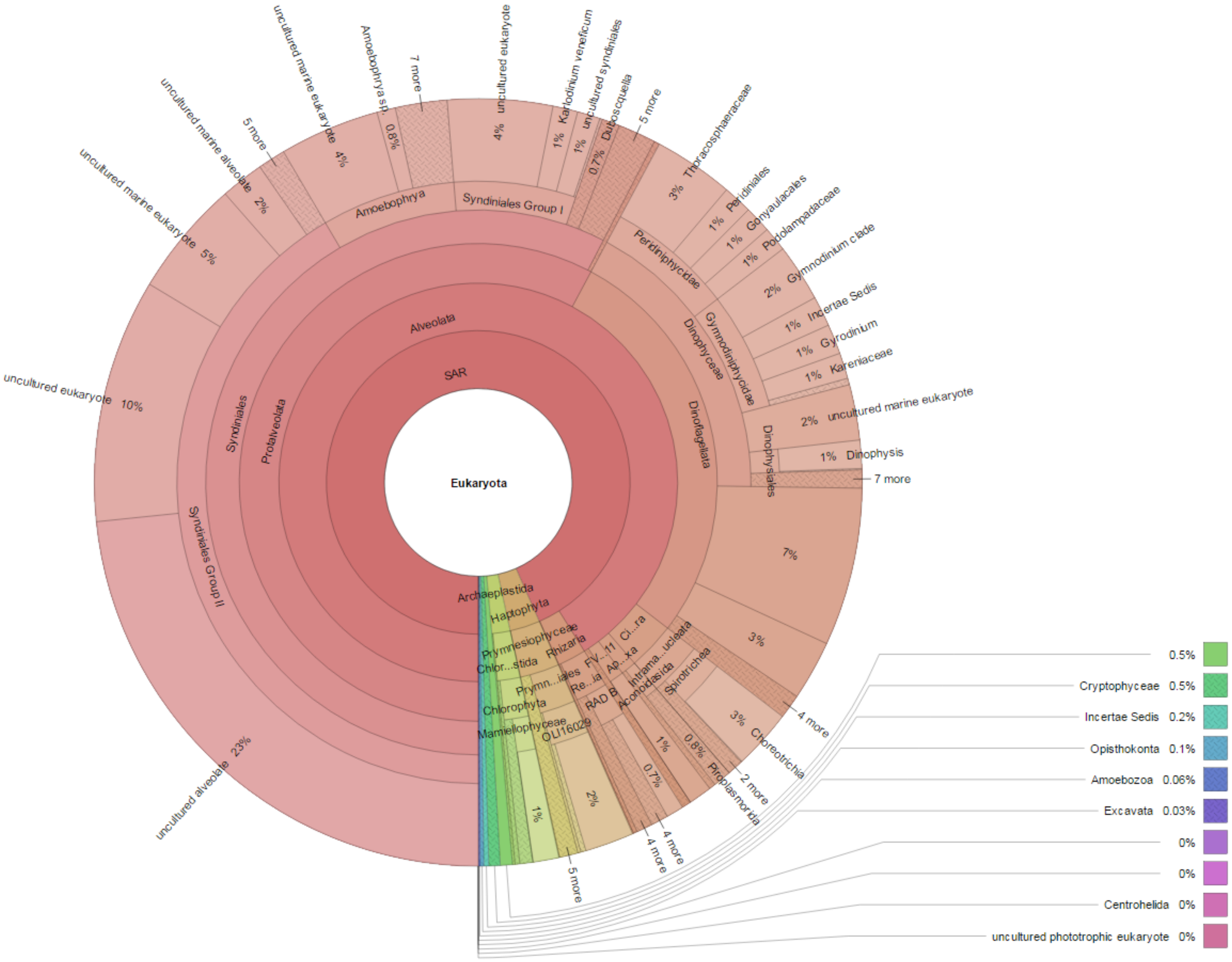
Supplementary Figure 13: Krona visualisation of the Eukaryotes in S3, showing average of the three replicates after removal of singletons.



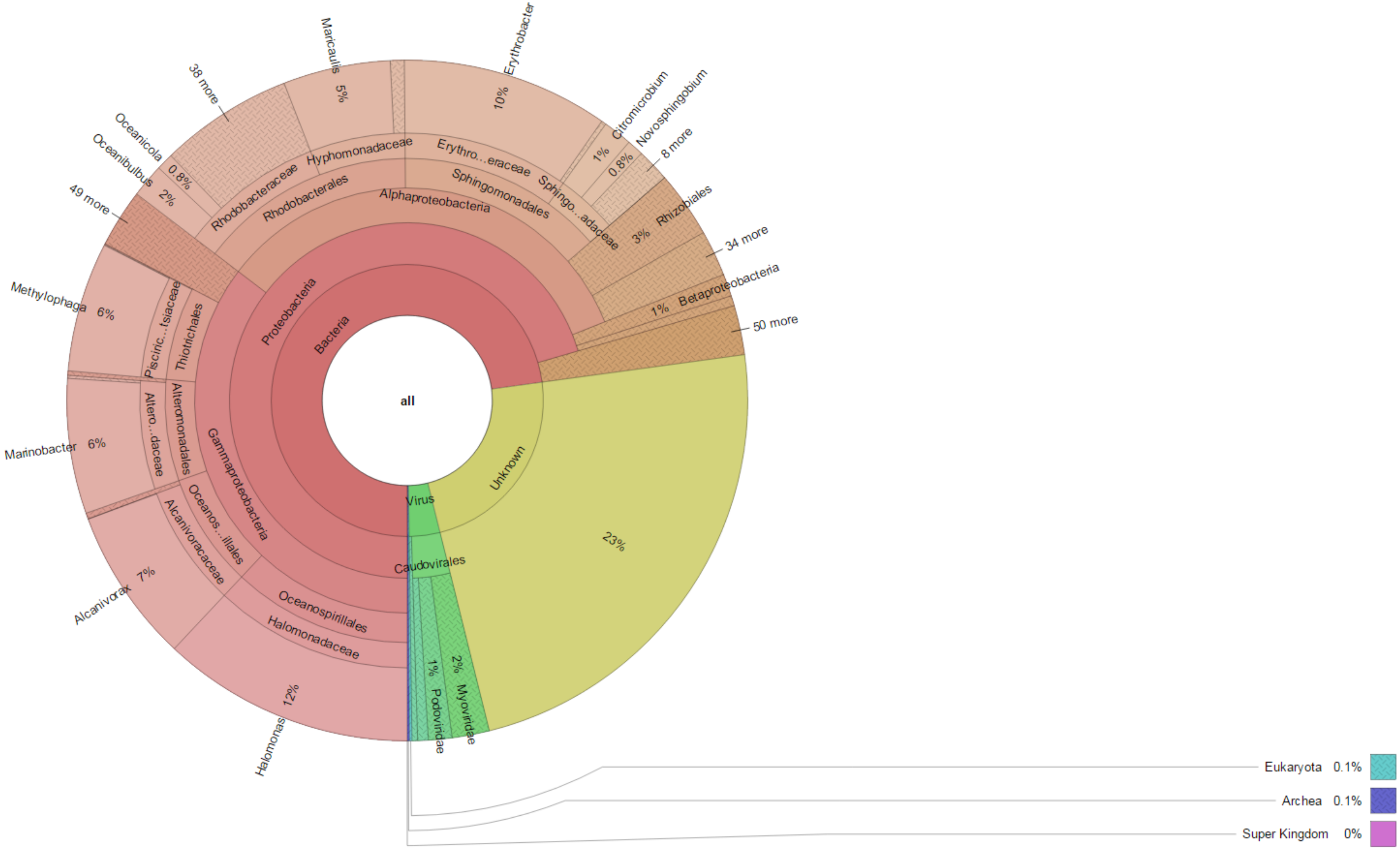
Supplementary Figure 14: Krona visualisation of the Eukaryotes in S4, showing average of the three replicates after removal of singletons.



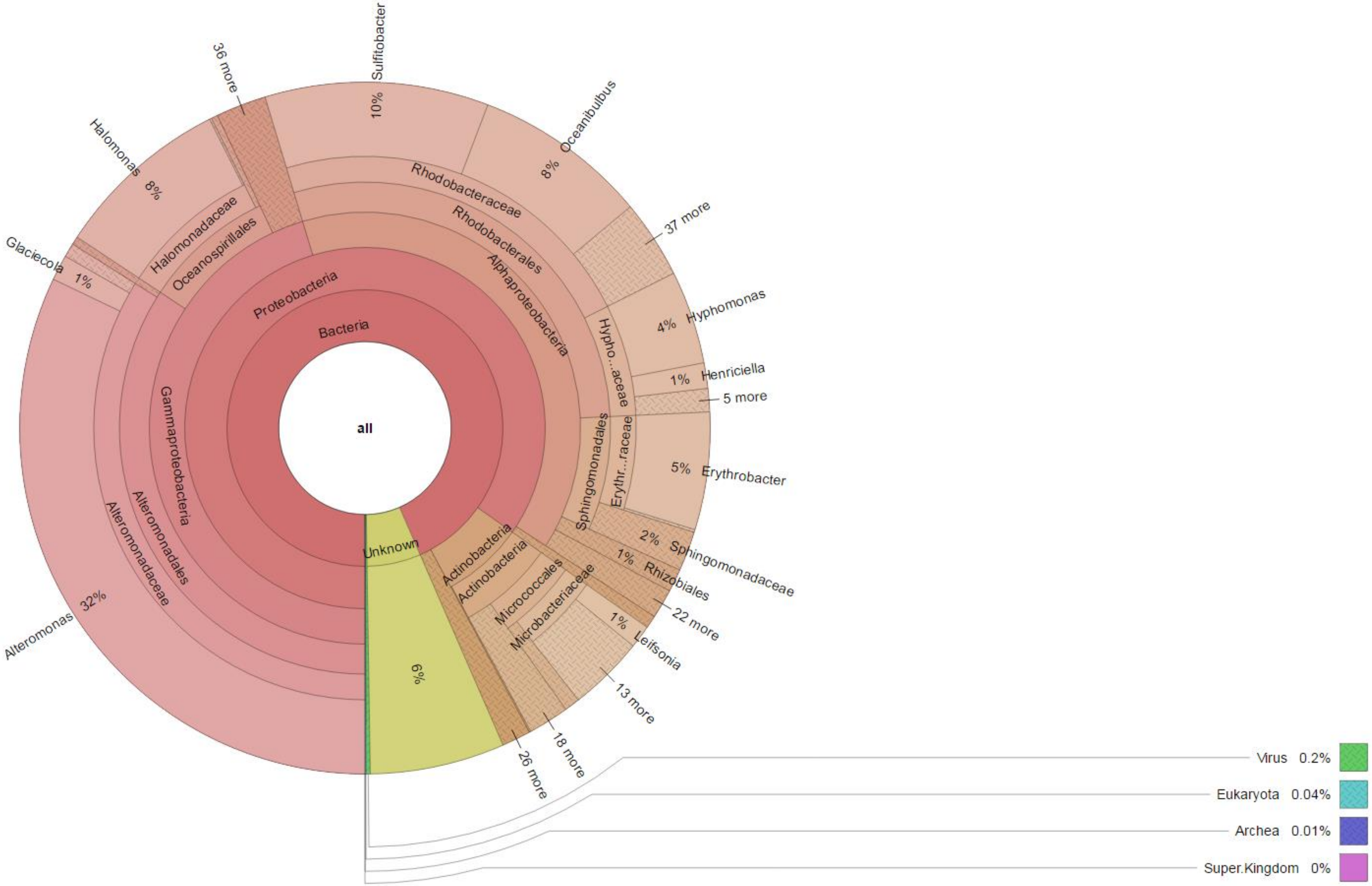
Supplementary Figure 15: Krona visualisation of the Eukaryotes in S5, showing average of the three replicates after removal of singletons.



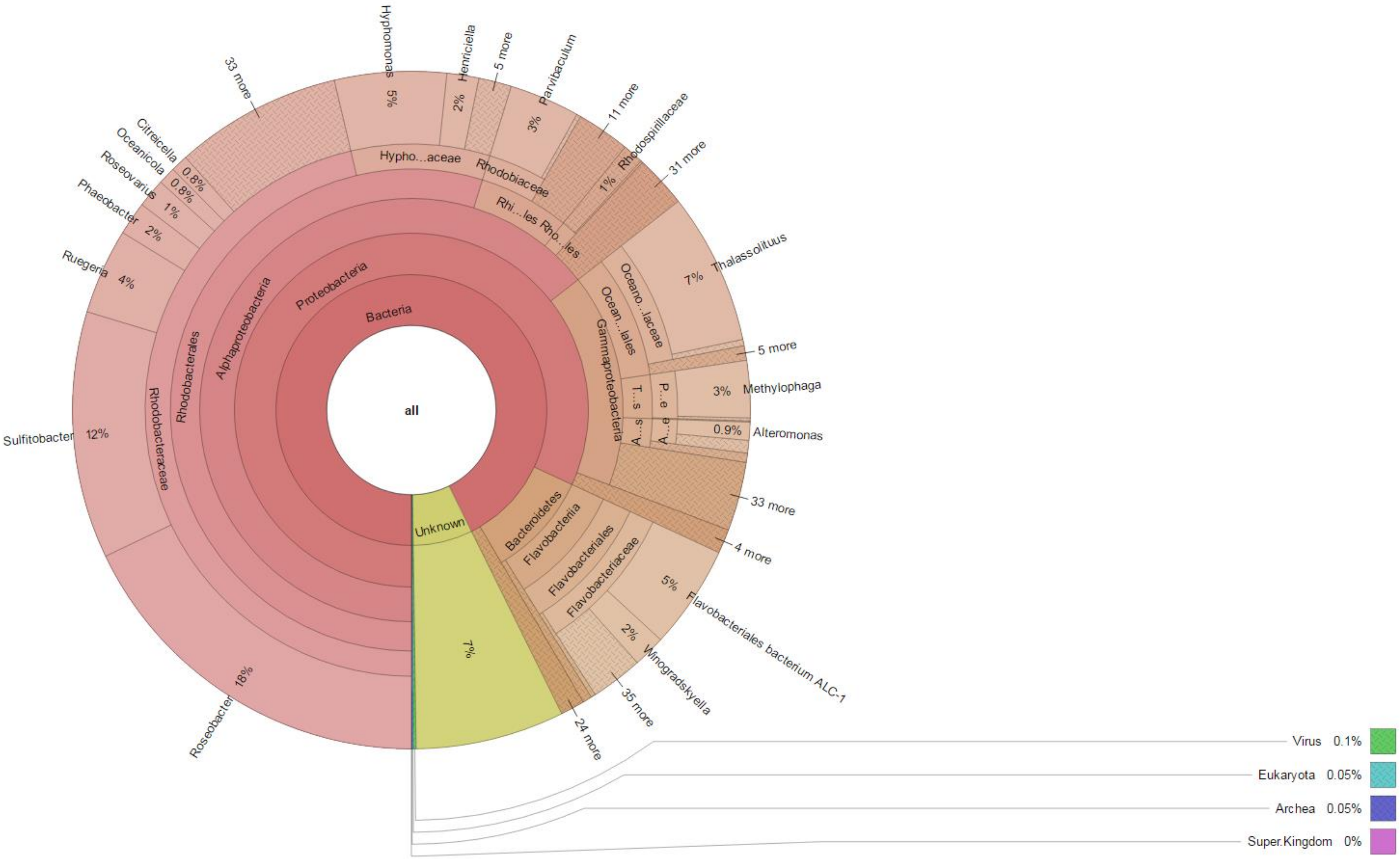
Supplementary Figure 16: Krona visualisation of the Eukaryotes in S6, showing average of the three replicates after removal of singletons.



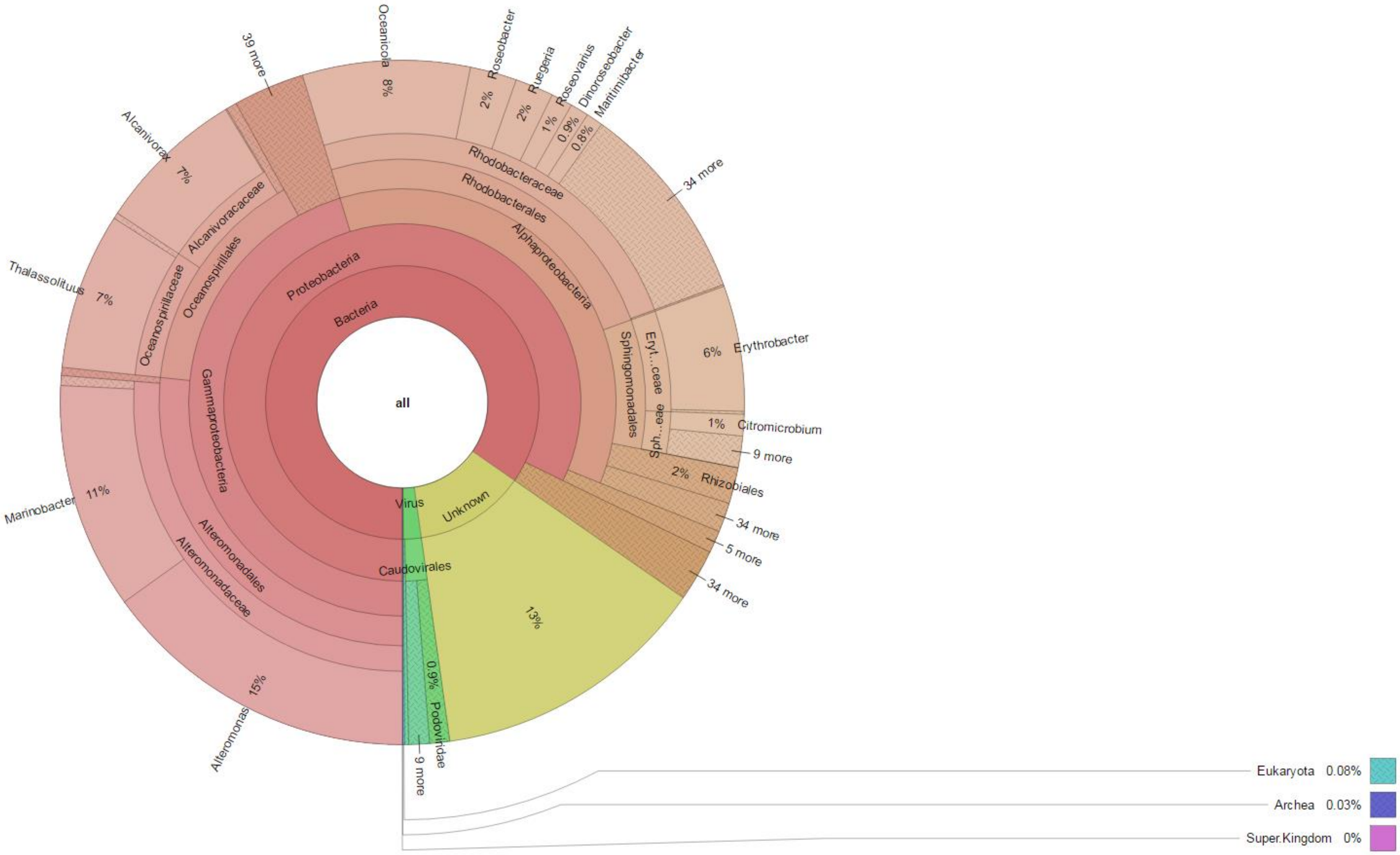
Supplementary Figure 17: Krona visualisation of the Refseq annotation of the permeate in S2



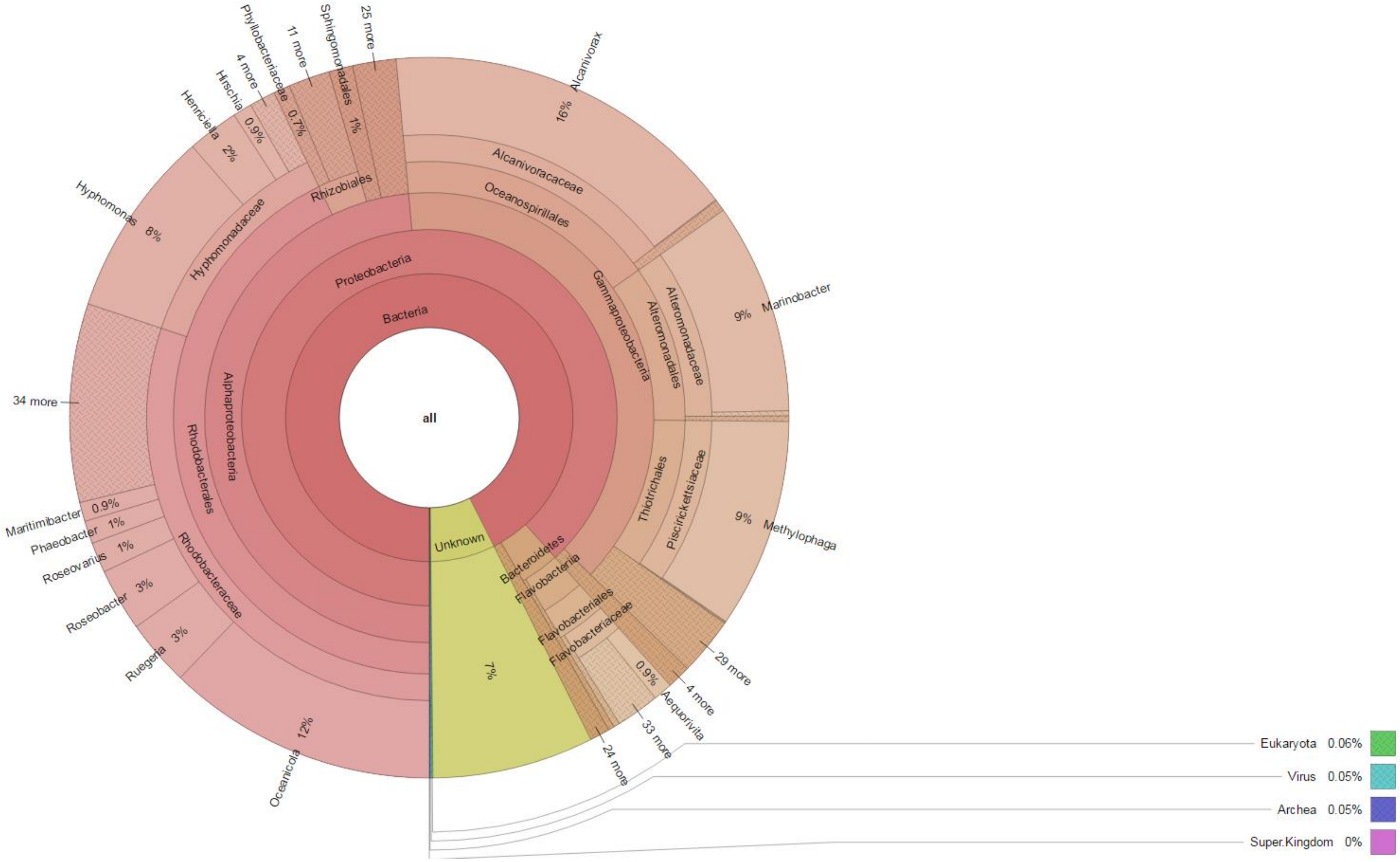
Supplementary Figure 18: Krona visualisation of the Refseq annotation of the permeate in S3



Supplementary Figure 19: Krona visualisation of the Refseq annotation of the permeate in S4



Supplementary Figure 20: Krona visualisation of the Refseq annotation of the permeate in S5



Supplementary Figure 21: Krona visualisation of the Refseq annotation of the permeate in S6

Supplementary Material: Tables

Supplementary Table 1: Information on sampling stations.

Station	**Area**	**Latitude**	**Longitude**	**Date**	**Depth**	**Temperature**
				(dd-mm-yyyy)	**(m)**	**(ºC)**
S1	South-West Indian Ocean	-38.314983	40.958083	22/02/2012	5	20.83
S2	South-West Indian Ocean	-35.507	37.4583	20/02/2012	49.09	19.98
S3	Southern Ocean	-57.5982	76.5083	06/03/2012	41.86	1.38
S4	Southern Ocean	-58.71	76.89	06/03/2012	40.93	1.24
S5	South-East Indian Ocean	-39.4753	108.9348	17/03/2012	44.978	16.23
S6	South-East Indian Ocean	-42.0817	113.3998	20/03/2012	60.55	12.95
S7	Elands Bay	-32.18618	18.19267	15/03/2013	<1	15.5
S8	Nelson Mandela Bay	-33.57086	25.38249	14/04/2013	<1	17.5
S9	Western Barents Sea	74.09	25.993	23/06/2012	20.2	5.89

Supplementary Table 2: Number of OTUs (97%) per sample following application of various data filters. T0p: no filter; T1: singleton removal; T5: filter removing OTUs observed with a total abundance <5; T10: filter removing OTUs observed with a total abundance <10. R1: filter retaining OTUs observed in at least two independent PCRs; R2: filter retaining OTUs observed in all three independent PCRs.

Filter	S1a	S1b	S2	S3	S4	S5	S6	S7	S8	S9
T0p	18215	12428	18203	15877	15603	22452	13675	15808	10486	14466
T1	13451	9812	13172	11093	11909	15543	9006	12177	8124	11404
R1	5283	2151	3703	2204	2882	4427	2803	3273	2413	2719
R2	4375	6727	5444	4204	5359	7891	4744	2323	3368	4469
T5	5323	4389	4961	4387	5011	6217	3169	5406	3629	4699
T10	2721	2351	2404	2209	2621	3153	1545	2929	2047	2517

Supplementary Table 3: Number of OTUs for the most abundant taxa based on samples and application of various filters. Mitochondria OTUs have been separated from the Alphaproteobacteria whilst Chloroplast have been separated from Cyanobacteria. S: station; Act: Actinobacteria; α: Alphaproteobacteria; Mit: mithocondria; Arc: Archaea; Bact: Bacteroidetes; β: Betaproteobacteria; Cyan: Cyanobacteria; Chl: chloroplast; δ: Deltaproteobacteria; γ: Gammaproteobacteria; NB: no blast hit; Plan: Planctomycetes; Ver: Verrucomicrobia.

S	Filter	Act	α	Mit	Arc	Bact	β	Cyan	Chl	δ	γ	NBt	Plan	Ver
S1a	No	274	4402	51	62	1306	119	2215	1290	632	2347	4672	48	222
	T1	197	3243	37	45	882	81	1629	904	491	1700	3665	31	149
	R1	56	1013	10	22	320	31	535	313	143	534	1213	7	56
	R2	112	1761	16	15	340	30	905	354	276	881	361	13	55
	T5	95	1254	16	10	237	28	671	315	238	629	1624	14	37
	T10	62	651	13	7	111	14	385	155	131	271	812	7	19
S1b	No	301	3412	42	39	1044	176	2078	1252	640	2378	152	64	235
	T1	257	2695	31	30	784	142	1760	929	509	1887	95	52	178
	R1	37	703	5	9	165	36	251	242	113	424	24	10	35
	R2	186	1775	17	20	533	73	1456	561	341	1303	46	23	121
	T5	112	1127	13	11	314	57	954	399	248	850	23	17	63
	T10	72	561	8	6	179	27	596	203	139	408	8	10	32
S2	No	571	3071	169	263	1274	92	3046	3269	1030	2413	1437	193	213
	T1	440	2236	127	206	910	54	2359	2339	744	1715	898	128	149
	R1	104	700	29	53	308	16	634	644	220	568	99	31	45
	R2	260	966	67	91	379	15	1118	1007	339	661	73	51	55
	T5	215	812	68	81	318	16	906	899	302	560	374	44	46
	T10	106	347	29	45	121	10	535	442	130	233	197	34	31
S3	No	17	1651	36	7	2333	303	81	3928	71	2057	5100	12	116
	T1	12	1142	21	5	1594	235	43	2906	41	1345	3540	10	91
	R1	2	306	5	2	474	88	16	744	6	355	121	1	29
	R2	4	651	7	3	868	111	16	1611	12	662	164	2	46
	T5	6	526	6	3	607	67	10	1251	15	572	1239	4	41
	T10	2	277	2	1	262	20	5	652	8	317	629	1	14
S4	No	19	3009	31	6	3181	430	19	4393	62	2348	1752	18	143
	T1	14	2289	23	4	2434	359	14	3400	46	1735	1315	12	123
	R1	2	567	6	1	657	91	5	859	10	471	146	0	37
	R2	4	1104	8	2	1185	182	3	1741	11	811	198	2	68
	T5	4	952	9	2	989	143	7	1511	20	681	578	4	61
	T10	2	582	3	1	448	37	4	786	8	386	323	1	19
S5	No	192	3925	136	137	2598	264	1066	5732	365	3511	3260	89	582
	T1	124	2596	95	89	1763	177	716	4191	250	2369	2286	54	415
	R1	51	820	30	29	554	44	162	1125	80	651	605	20	138
	R2	57	1457	50	53	947	110	469	2477	109	1404	291	26	230
	T5	26	1006	30	23	578	85	356	1771	79	1027	916	14	134
	T10	9	538	19	10	258	58	232	855	42	516	482	8	56
S6	No	173	2189	77	215	1699	225	739	4357	258	2409	305	101	333
	T1	118	1435	58	153	1049	141	529	3016	168	1517	181	55	208
	R1	28	439	15	49	380	33	135	959	46	459	61	21	63
	R2	74	813	21	81	515	80	340	1510	73	846	64	20	119
	T5	50	547	17	46	278	56	250	1001	64	569	59	16	78

	T10	17	265	9	18	132	19	126	494	28	312	25	9	36
S7	No	80	1712	111	98	1873	129	201	2132	166	4377	4296	92	133
	T1	57	1279	87	73	1409	98	156	1645	110	3261	3435	66	105
	R1	3	334	1	2	319	6	55	558	10	863	993	16	39
	R2	5	257	0	1	127	3	52	354	4	1053	369	7	7
	T5	20	611	30	29	513	34	79	708	42	1381	1708	20	45
	T10	18	330	9	11	239	14	43	341	27	749	986	13	22
S8	No	113	1591	123	40	4053	225	156	658	71	1802	1247	62	50
	T1	96	1198	96	31	3159	181	136	490	37	1276	1062	47	41
	R1	33	314	20	4	897	47	27	142	5	360	487	19	10
	R2	27	537	57	7	1600	91	87	222	6	556	48	5	7
	T5	43	517	57	13	1389	81	80	197	10	556	491	26	21
	T10	22	304	41	7	718	62	32	116	7	305	286	19	16
S9	No	86	2332	81	82	2386	253	27	2464	58	2981	3457	23	56
	T1	72	1886	70	60	1930	193	22	1961	42	2392	2567	19	46
	R1	23	502	18	14	494	67	2	488	8	577	478	0	16
	R2	33	920	38	19	962	67	6	1005	10	1245	98	4	15
	T5	21	747	36	16	787	53	6	802	16	1009	1141	6	16
	T10	7	382	15	8	438	18	5	426	9	573	608	2	8

Supplementary Table 4: List of phylotypes present only at T0.

18s species presents only at T0	**16s species presents only at T0**				
Acanthamoeba castellanii	Acanthopleuribacter pedis	Cucumis sativus	Leucothrix	Roseobacter sp. QSSC9-8	uncultured Oceanibaculum sp.
Amphibelone anomala	Achromatium minus	Curvibacter	Liberibacter crescens BT-1	Roseobacter sp. SDT2S7	uncultured Oscillatoriales cyanobacterium
Amphidinium belauense	Achromatium oxaliferum	Cyanidium caldarium	Limibacter	Roseomonas	uncultured Paracoccus sp.
Amphora cf. proteus	Achromatium sp. JD1	cyanobacterium SC-1	Limnobacter	Rubidimonas	uncultured Pelobacter sp.
Andalucia incarcerata	Achromobacter	Cytophaga-like bacterium QSSC1-18	Limnothrix	Sagittula	uncultured Photobacterium sp.
Ankistrodesmus sp. Mary 8/18 T-2w	Acidimicrobiaceae	Cytophaga sp. I-545	Lishizhenia tianjinensis	Salinibacter	uncultured Pirellula sp.
Anurofeca sp. LAH-2003	Acidiphilium	Deinococcus	Loktanella salsilacus	Salinisphaera	uncultured Piscirickettsia sp.
Apatococcus lobatus	Acidobacteriaceae	Delphineis sp. CCMP1095	LPP-group cyanobacterium QSSC5cya	Salinisphaera orenii MK-B5	uncultured Planctomyces sp.
Apicomplexa	Acidobacterium	delta proteobacterium enrichment culture clone VNABa05	Luteibacter	Salinisphaeraceae	uncultured Planctomycetaceae bacterium
Aureococcus anophagefferens	Actibacterium mucosum	delta proteobacterium PSCGC 5342	Magnetovibrio	Sandaracinobacter	uncultured Pleurocapsa sp.
Beroe ovata	Actinobacillus	denitrifying bacterium enrichment culture clone NOB_2_C8	Marine Methylotrophic Group 2	Sandaracinus	uncultured Polyangiaceae bacterium
Betula platyphylla	Actinobacteria bacterium canine oral taxon 376	Dermacoccus nishinomiyaensis	Marine Methylotrophic Group 3	Sandarakinorhabdus	uncultured Prochloron sp.
Blastocystis sp. C12	Actinomyces	Desulfarculus	Marinobacter lutaoensis	Schleiferia	uncultured Pseudoalteromonas sp.
Blastodinium mangini	Aeromonas sp. RR8	Desulfatirhabdium	Marinobacterium jannaschii	secondary endosymbiont of Amonostherium lichtensioides	uncultured Pseudomonadales bacterium

Blastodinium sp. 2 CAdS-2011
Botryococcus sp. UTEX 2629
Brassica rapa
Bucegia romanica
Cavostelium apophysatum
Chaetoceros
Chloromonas insignis
Chlorophyta sp. CCMP1407
Chrysochromulina campanulifera
Chrysophyceae
Cicer arietinum
Coccolithus pelagicus
Coccomyxa parasitica
Cochlodinium
Collozoum pelagicum
Cosmarium protractum

Aeromonas veronii
Agaricicola
Agarivorans albus
Agrobacterium tumefaciens
Ahrensia kielensis
Alexandrium tamarense
Algimonas
Algimonas porphyrae
Alkaliphilus
alpha proteobacterium HTA473
alpha proteobacterium MBIC3035
alpha proteobacterium ML-126
alpha proteobacterium MN-5
Anaerococcus
Anaplasmataceae
Ancalomicrobium

Desulfobacca
Desulfobacterium
Desulfobacterium indolicum
Desulfocella
Desulfococcus
Desulfococcus multivorans
Desulfohalobiaceae
Desulfomicrobium orale
Desulfomicrobium sp. enrichment culture clone LDC-15
Desulfomonile
Desulfonatronobacter
Desulfonatronovibrio thiodismutans
Desulfonatronum thioautotrophicum
Desulfonema ishimotonii
Desulforegula
Desulforhabdus

Marinovum
Mariprofundaceae
Mariprofundus
Marivita
Marixanthomonas
Meganema
Merismopedia punctata PMC242.05
Mesonia
Mesorhizobium ciceri
Methylocaldum
Methylococcaceae
Methylocystis
Methylohalobius
Methylomicrobium agile
Methylorosula
Methylosinus trichosporium

secondary endosymbiont of Aphalaroida inermis
secondary endosymbiont of Calophya schini
Sedimentibacter
Selenomonas
Serratia
Serratia symbiotica str. Tucson
Shewanella sp. AK55
Shewanella sp. enrichment culture clone PKWE30-13
Shewanella sp. KJF13-1
Sinorhizobium sp. JNVU AN2
Skermanella
Sphingomonadaceae
Sphingomonas sp. JS8(2011)
Sphingopyxis sp. 14C-7
Sphingopyxis sp. BB24
Spiribacter salinus M19-40

uncultured Ralstonia sp.
uncultured Rheinheimera sp.
uncultured Rhodospirillales bacterium
uncultured SAR156 cluster gamma proteobacterium
uncultured Sediminibacterium sp.
uncultured Shewanella sp.
uncultured Sodalis sp.
uncultured spirochete
uncultured sulfur-oxidizing symbiont bacterium
uncultured Sulfurovum sp.
uncultured SUP05 cluster bacterium
uncultured Syntrophaceae bacterium
uncultured Syntrophobacter sp.
uncultured Thalassobius sp.
uncultured Thalassolituus sp.
uncultured Thermoactinomyces sp.

Cryptophyta sp. SL64/78sp
Cryptosporidium sp.
Cyclotella
Desmarestia viridis
Diaphanoeca
Diaphanoeca grandis
Diatoma cf. tenuis
Dictyocha speculum
Dinobryon sertularia
Dixoniella grisea
Dothideales
Dunaliella bardawil
Echinamoeba
Fabomonas tropica
Flintiella sanguinaria
Fragilariales
Freshwater Choanoflagellates 1
Gracilariopsis chorda
Gromia
Guillardia theta

Anderseniella
Anoxybacillus
Antithamnion sp.
Arctic sea ice bacterium ARK10036
Arsenophonus
Arthrobacter sp. L-6
Aspergillus clavatoflavus
Azoarcus
Azomonas insignis
Azospira
Bacillales
Bacillus azotoformans LMG 9581
Bacillus megaterium
Bacillus sp. 7-8
Bacillus sp. B10 ZZ-2008
bacterium 20N1
bacterium BW3PhS19
bacterium DG1021
bacterium DG1026
bacterium EJ10-97

Desulfosporosinus
Desulfothermus naphthae
Desulfovibrionaceae
Desulfovirga
Desulfurella
Desulfurivibrio
Desulfuromusa
Devosia
Diaphorobacter
Dichotomicrobium
Dinoroseobacter shibae
Dokdonia
Donghicola
Ectothiorhodospira sp. 'Bogoria Red'
Elusimicrobia
endosymbiont of Acanthamoeba sp. UWC36
endosymbiont of Columbicola baculoides
endosymbiont of Columbicola macrourae
Enhygromyxa salina
Enterobacter sp. NCCP-195

Microbacterium
Microbacterium soli
Microbacterium sp. YRR08
Microbulbifer maritimus
Micrococcus
Micrococcus sp. VKRKCo13
Microcystis
Microcystis elabens
Micromonospora
Moorea producens NAC8-48
Moraxella sp. MOR44
Muricauda
Mycoplasma
Mycoplasma aquilae ATCC BAA-1896
Mycoplasma oxoniensis
Nannochloropsis oceanica
Nautilia
Neisseriaceae
Neorickettsia
Nesterenkonia sp. DSM 27373

Spirulina subsalsa
Spirulina subsalsa IAM M-223
Sporichthya
Stakelama sp. JC126
Stella
Steroidobacter
Streptomyces sp. AV050
Streptomyces sp. CLS45
Sulfobacillus
Sulfuricella
Sunxiuqinia
Sutterella
Synechococcus sp. MBIC10613
Synechococcus sp. MW10#1
Synergistales
Syntrophaceae
Syntrophobacter
Tabrizicola
Taibaiella
Terrabacter

uncultured Thermus/Deinococcus group bacterium
uncultured Thiomicrospira sp.
uncultured Tistrella sp.
uncultured Ulvibacter sp.
uncultured Verrucomicrobium sp.
Variovorax paradoxus
Verminephrobacter
Vibrio aestuarianus
Vibrio agarivorans
Vibrio campbellii
Vibrio coralliilyticus
Vibrio marisflavi CECT 7928
Vibrio mytili
Vibrio parahaemolyticus
Vibrio sp. 0208F2
Vibrio sp. 3d
Vogesella
Woodsholea
Xanthobacillum maris
Xanthomonas

Haliclona sp. OGL2003
Halimeda renschii
Haliommatidium sp.
Halostylodinium
Haptolina brevifila
Hedychium
Hemiselmis brunnescens
Hepatozoon sp. Boiga
Histioneis sp. FTL62
Holmsella pachyderma
Kentrophoros gracilis
Kupea martinetugei
Leucolepis menziesii
Lotus japonicus
Marine Choanoflagellates 1
Microporella ciliata

bacterium Ellin5257
bacterium endosymbiont of Ischnodemus sabuleti
bacterium endosymbiont of Lipoptena depressa
bacterium endosymbiont of Osedax mucofloris
bacterium enrichment culture clone BBMC-4
bacterium enrichment culture clone CBNH_1102_HA1_BOTTOM_10
bacterium enrichment culture clone EB27.11
bacterium enrichment culture clone EtOH-57
bacterium enrichment culture clone LA29
bacterium enrichment culture clone R4-32B
bacterium enrichment culture clone SBII3
bacterium enrichment culture clone Tol_7
bacterium GLA1
bacterium IS6
bacterium SH5-11
Bacteroidetes bacterium CNX-216

Enterobacteriaceae
Enterobacteriaceae bacterium secondary endosymbiont of Crisicoccus azaleae
enterobacterium dtb112
Enterovibrio nigricans
Erwinia
Erythrobacteraceae
Eubacterium infirmum
Euglena gracilis
Euglena proxima
Euptilota molle
Euzebya
Ferrimonas sp. EF3B-B688
Ferriphaselus
Fibrobacterales
Fibrobacteria
Filomicrobium

Nicotiana sylvestris
Nitriliruptoraceae
Nitrosomonas
Nitrospirales
Oceaniserpentilla
Oceanospirillaceae
Olavius loisae endosymbiont 3
Oleispira lenta
Oligosphaeria
Orenia sivashensis
Ornatilinea
Oscillatoria sp. LEGE 05292
Oscillatoria spongeliae SI04-46
Ottowia
Paenibacillaceae
Paenibacillus sp. WP7

Terrestrial Miscellaneous Gp(TMEG)
Terrimonas
Tetraselmis cordiformis
Thalassomonas loyana
Thalassomonas sp. CL-22
Thalassomonas sp. PaD1.04
Thiofaba
Thiohalobacter
Thiohalospira alkaliphila
Tissierella
Truepera
Tsukamurella paurometabola
uncultured Acidimicrobidae bacterium
uncultured Acidimicrobineae bacterium
uncultured Acidobacteriaceae bacterium
uncultured Acidobacterium sp.

Xenorhabdus japonica
Zooshikella ganghwensis DSM 15267
Zymomonas mobilis subsp. francensis
Zymomonas mobilis subsp. mobilis str. CP4 = NRRL B-14023

Mnium hornum
Monomastix minuta
Monosporus pedicellatus
Musa acuminata subsp. malaccensis
Nematoda environmental sample
Neorhodella cyanea
Nicotiana sylvestris
Nitzschia communis
Oblongichytrium sp. HK9
Paragordionus dispar
Pelagococcus subviridis
Penium margaritaceum
Peridinium willei
Pfiesteria-like sp. F525Jul02
Phaeodactylum tricornutum
Phagocata ullala
Phaseolus vulgaris

Bangiopsis subsimplex
Bartonella
Bathymodiolus brooksi gill symbiont
Beggiatoa
Beijerinckiaceae
benzene mineralizing consortium clone SB-30
Bisgaard Taxon 37
Blastocatella
Blochmannia endosymbiont of Opisthopsis haddoni 244
Bowmanella pacifica
Bowmanella sp. UDC354
Brachybacterium
Brumimicrobium
Buchnera aphidicola (Takecallis arundicolens)
Calothrix
Campylobacter
Candidate division BRC1

Flammeovirgaceae
Flavobacteria bacterium CC-AMO-30D
Flavobacteriaceae bacterium Hel_I_10
Flavobacteriaceae bacterium JBKA-6
Flavobacteriaceae bacterium LPK5
Flavobacterium columnare
Flavobacterium enshiense DK69
Flavobacterium sp. HME6133
Flavobacterium sp. WB 3.1-78
Flavobacterium sp. WB3.4-76
Fonticella tunisiensis
Francisella sp. 10HP457
Frigoribacterium
Gaiellales
Galbibacter sp. NBRC 101636
Gallionella
gamma proteobacterium endosymbiont of Pseudococcus viburni

Pannonibacter phragmitetus
Parabacteroides
Paracoccus-like sp. V4.BP.10
Paraliobacillus ryukyuensis
Parvularcula
Patulibacter
Pediococcus pentosaceus
Pelagibacterium
Perlucidibaca
Persicobacter sp. JZB09
Phaeodactylum tricornutum
Planktothrix sp. PCC 9214
Polaribacter sp. SM1202
Polynucleobacter
Ponticaulis koreensis DSM 19734
Prasinococcus capsulatus
Pricia

uncultured Acinetobacter sp.
uncultured Aeromonas sp.
uncultured Alicyclobacillaceae bacterium
uncultured Antarctic sea ice bacterium
uncultured Arcobacter sp.
uncultured Azomonas sp.
uncultured Bacteriovorax sp.
uncultured bacterium gp1
uncultured bacterium HERMI11
uncultured bacterium HF0770_08F21
uncultured bacterium KM3-23-D4
uncultured bacterium KM3-69-D9
uncultured bacterium tbr1-3
uncultured Banisveld landfill bacterium BVB22
uncultured Beggiatoa sp.
uncultured Bradyrhizobiaceae bacterium
uncultured Caldilinea sp.

Pinus taeda
Porphyridium aerugineum
Prorocentrum consutum
Prorocentrum glenanicum
Prorocentrum rhathymum
Protaspa
Protaspis sp. CC-2009c
Pseudo-nitzschia cuspidata
Pseudo-nitzschia seriata
Pseudopedinella
Psilopilum laevigatum
Pyrenomonas salina
Pyrosoma atlanticum
Rheomorpha neiswestonovae
Rhizamoeba saxonica
Riccia huebeneriana subsp. sullivantii

Candidate division SR1
Candidate division WS3
Candidatus Accumulibacter
Candidatus Anammoxoglobus propionicus
Candidatus Ancillula trichonymphae
Candidatus Blochmannia
Candidatus Blochmannia rufipes
Candidatus Branchiomonas
Candidatus Cryptoprodotis polytropus
Candidatus Curculioniphilus buchneri
Candidatus Ecksteinia adelgidicola
Candidatus Endoecteinascidia frumentensis
Candidatus Marithrix
Candidatus Methylomirabilis
Candidatus Photodesmus katoptron Akat1
Candidatus Profftia virida

gamma proteobacterium enrichment culture clone DT-1983
gamma proteobacterium MOLA 531
Gangjinia
Gangjinia marincola
Gemella
Gemmatimonadaceae
Gemmatimonadetes
Geobacillus
Gilvibacter
Gilvimarinus
Glaciecola sp. Za3-19
Gloeocalita
Granulicella
Guillardia theta
Haemophilus
Halalkalicoccus paucihalophilus

primary endosymbiont of Pseudolynchia canariensis
Prochlorothrix hollandica PCC 9006
Propionibacterium
Proteiniborus
Pseudendoclonium akinetum
Pseudoalteromonas espejiana
Pseudoalteromonas sp. avm16
Pseudoalteromonas sp. BCw029
Pseudoalteromonas sp. BSi20680
Pseudoalteromonas sp. BSw20514
Pseudoalteromonas sp. BSw21454
Pseudoalteromonas sp. DS-12
Pseudoalteromonas sp. WZUC10
Pseudoalteromonas tunicata
Pseudomonas guineae
Pseudomonas poae

uncultured candidate division OS-K bacterium
uncultured candidate division TG3 bacterium
uncultured candidate division WS5 bacterium
uncultured Caulobacterales bacterium
uncultured Cellvibrio sp.
uncultured Chlorobiales bacterium
uncultured Chloroflexaceae bacterium
uncultured Clostridiaceae bacterium
uncultured Clostridiales bacterium
uncultured Comamonadaceae bacterium
uncultured compost bacterium
uncultured Crocinitomix sp.
uncultured Dechloromarinus sp.
uncultured Deinococcales bacterium
uncultured Delftia sp.
uncultured Desulfobulbus sp.

Saccamoeba limax
Savillea micropora
Scenedesmaceae sp. Tow 9/21 P-14w
Schizoplasmodium cavostelioides
Scrippsiella sp. SCKS 0701
Sesamum indicum
Sinophysis stenosoma
Sitodiplosis mosellana
Skeletonema grevillei
Sorghum bicolor
Spumellaria
Symbiodinium pilosum
Tabularia tabulata
Talaromyces purpurogenus
Telonema

Candidatus Purcelliella pentastirinorum
Candidatus Stammerula sp. of Acanthiophilus helianthi
Candidatus Thiobios
Candidatus Zinderia insecticola
Capnocytophaga cynodegmi
Catenovulum
Celerinatantimonas
Celerinatantimonas diazotrophica
Cerasicoccus
cf. Wilmottia sp. CAWBG522
Chitinophaga
Chlamydiales symbiont of Salmo salar
Chlorella mirabilis
Chlorobiaceae
Chloroflexia

Haliangium tepidum
Halochromatium
Halomonadaceae bacterium PH27A
Haloplasma
Halothiobacillus
Halyomorpha halys symbiont
Helicobacteraceae
Hellea
Holophaga
Holophagaceae
Hyphomonadaceae bacterium JC2236
Hypnea sp. DWF-2004
Ideonella
Ilumatobacter
iron-reducing bacterium enrichment culture clone HN-HFO22

Pseudomonas sp. CB-11
Pseudomonas sp. HC1-18
Pseudomonas sp. PHAs045
Pseudoruegeria
Pseudoscourfieldia sp. Nak
Pseudoteredinibacter
Psychromonas
psychrophilic sulfate-reducing bacterium LSv55
Psychroserpens
Puniceicoccus
Pyramimonas tetrarhynchus
Reinekea
Rhizobiales Incertae Sedis
Rhodanobacter
Rhodobacter sphaeroides

uncultured Desulfomicrobium sp.
uncultured Desulfovibrionaceae bacterium
uncultured Devosia sp.
uncultured Ectothiorhodospiraceae bacterium
uncultured Enterococcus sp.
uncultured eubacterium CHA3-117
uncultured Ferruginibacter sp.
uncultured Fibrobacteres/Acidobacteria group bacterium
uncultured Flammeovirgaceae bacterium
uncultured Flexibacteraceae bacterium
uncultured Fluviicola sp.
uncultured forest soil bacterium
uncultured gamma proteobacterium CHAB-IV-19
uncultured gamma proteobacterium HF0200_24F15
uncultured gamma proteobacterium HF4000_19M20

Thalassionema bacillare	Christensenella	Janthinobacterium sp. 42	Rhodobacteraceae bacterium D5-6.1	uncultured Granulosicoccaceae bacterium
Thalassiosira punctigera	Chroococcus	Kangiella	Rhodobiaceae	uncultured Haliscomenobacter sp.
Thalassiosira weissflogii	Chungangia	Kistimonas	Rhodocyclaceae	uncultured hydrocarbon seep bacterium BPC065
Thecofilosea	Citrobacter	Klebsiella	Rhodomicrobium	uncultured Hydrogenophaga sp.
Theileria annulata	Clostridiisalibacter	Kryptophanaron alfredi symbiont	Rhodomonas salina	uncultured Hyphomicrobiaceae bacterium
Trebouxiophyceae	Clostridium sensu stricto 1	Lactococcus lactis subsp. lactis	Rhodospira trueperi	uncultured Lactobacillus sp.
Trichodina pectenis	Coccinimonas marina	LD28 freshwater group	Rhodothermaceae	uncultured low G+C Gram-positive bacterium
Trimastix marina	Cohaesibacter	Legionella sp. LC2720	Rhodovulum phaeolacus	uncultured Lutibacter sp.
uncultured Boletaceae	Collinsella	Legionella tucsonensis	Rhynchosporium agropyri	uncultured Lysobacter sp.
uncultured Chlamydomonadaceae	Colwelliaceae	Legionellaceae	Rickettsia asiatica	uncultured Magnetococcus sp.
uncultured Closteriaceae	Compsopogon caeruleus	Leptolyngbya sp. CCAP 1442/1	Rickettsiaceae bacterium Os18	uncultured marine group I thaumarchaeote
uncultured Oxytrichidae	Conchiformibius	Leptolyngbya sp. LEGE 07088	Rivularia sp. PCC 7116	uncultured Marinobacter sp.
uncultured phototrophic eukaryote	Corynebacterium	Leptolyngbya sp. OU_8	Robiginitomaculum antarcticum DSM 21748	uncultured methanogenic archaeon
uncultured Telonema	Coxiellaceae	Leptolyngbya sp. RS03	Roseibacterium sp. JLT1202r	uncultured Methylococcaceae bacterium
unidentified protist 56059	Crenothrix	Leptospirillum	Roseobacter clade CHAB-I-5 lineage	uncultured Microgenomates bacterium
Urospora neglecta	Criblamydia sequanensis	Leptotrichiaceae	Roseobacter clade Marinomonas lineage	uncultured Novosphingobium sp.

Warnowia sp. BSL-2009a	Cryomorpha

Supplementary Table 5: Presence - absence comparison. Comparison across thresholds from T0 to T10-R2 applied on both amplicon dataset and the metagenomics dataset unfiltered (T0).

	All contigs;	All contigs;	All contigs;	All contigs;	All contigs;
Common	Phaeodactylum				
Common 16s- 18s	Alexandrium				
	Chlorella	Chlorella			
	Chroomonas	Chroomonas			
	Cicer				
	Cucumis				
	Cymbomonas	Cymbomonas			
	Flintiella				
	Guillardia				
	Karlodinium	Karlodinium			
	Nicotiana				
	Pedinomonas	Pedinomonas			
	Phaeodactylum				
	Picea	Picea	Picea	Picea	
	Porphyridium				
	Prasinoderma	Prasinoderma	Prasinoderma		
	Prototheca	Prototheca			
	Prymnesium	Prymnesium	Prymnesium	Prymnesium	Prymnesium
	Pyramimonas				
Common metagenome - 18s	Desmarestia				
	Phaeodactylum				
Common metagenome - 16s	Achromobacter				
	Acidovorax	Acidovorax	Acidovorax	Acidovorax	Acidovorax
	Acinetobacter	Acinetobacter			
	Actinobacteria	Actinobacteria	Actinobacteria	Actinobacteria	Actinobacteria
	Actinomyces	Actinomyces	Actinomyces	Actinomyces	Actinomyces
	Afipia				
	Agrobacterium				

	Ahrensia				
	Alcanivorax	Alcanivorax	Alcanivorax		
	Algicola	Algicola			
	Alicyclobacillus				
	Alkaliphilus				
	Alphaproteobacteria	Alphaproteobacteria	Alphaproteobacteria	Alphaproteobacteria	Alphaproteobacteria
	Alteromonas	Alteromonas	Alteromonas	Alteromonas	Alteromonas
	Arthrobacter	Arthrobacter	Arthrobacter	Arthrobacter	
	Asticcacaulis	Asticcacaulis			
	Azoarcus				
	Azotobacter				
	Bacillales				
	Bacillus	Bacillus	Bacillus	Bacillus	
	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
	Betaproteobacteria	Betaproteobacteria	Betaproteobacteria	Betaproteobacteria	Betaproteobacteria
	Blastomonas				
	Brachybacterium				
	Bradyrhizobium	Bradyrhizobium	Bradyrhizobium		
	Brevundimonas	Brevundimonas			
	Burkholderia	Burkholderia	Burkholderia	Burkholderia	Burkholderia
	Calothrix	Calothrix			
	Campylobacter				
	Candidatus Aquiluna	Candidatus Aquiluna			
	Candidatus Liberibacter				
	Celeribacter				
	Chitinophaga				
	Chroococcidiopsis	Chroococcidiopsis			
	Citrobacter				
	Clostridium				
	Corynebacterium				
	Cupriavidus	Cupriavidus	Cupriavidus		

	Curvibacter				
	Cyanothece	Cyanothece			
	Cycloclasticus	Cycloclasticus			
	Dechloromonas	Dechloromonas			
	Deinococcus				
	Desulfotomaculum	Desulfotomaculum			
	Dinoroseobacter				
	Enterococcus				
	Erythrobacter	Erythrobacter	Erythrobacter	Erythrobacter	Erythrobacter
	Erythrobacteraceae				
	Fibrella	Fibrella			
	Flavobacterium	Flavobacterium	Flavobacterium	Flavobacterium	Flavobacterium
	Fodinicurvata				
	Gammaproteobacteria	Gammaproteobacteria	Gammaproteobacteria	Gammaproteobacteria	Gammaproteobacteria
	Gemmata	Gemmata	Gemmata		
	Gilvimarinus				
	Glaciecola	Glaciecola			
	Gracilimonas	Gracilimonas			
	Haliangium				
	Halomonas	Halomonas	Halomonas		
	Hirschia	Hirschia			
	Hoeflea	Hoeflea	Hoeflea	Hoeflea	
	Hydrogenophaga	Hydrogenophaga			
	Hyphomicrobium	Hyphomicrobium	Hyphomicrobium		
	Hyphomonas	Hyphomonas	Hyphomonas		
	Ideonella				
	Ilumatobacter				
	Janthinobacterium				
	Kiloniella	Kiloniella			
	Kordiimonas	Kordiimonas			
	Lachnospiraceae	Lachnospiraceae			

	Legionella	Legionella			
	Lentisphaera	Lentisphaera			
	Leucothrix				
	Limnobacter				
	Loktanella	Loktanella	Loktanella	Loktanella	Loktanella
	Magnetospirillum	Magnetospirillum			
	Marinobacter	Marinobacter			
	Marinobacterium	Marinobacterium	Marinobacterium	Marinobacterium	
	Massilia	Massilia	Massilia	Massilia	Massilia
	Mesorhizobium	Mesorhizobium			
	Methylobacterium	Methylobacterium			
	Methylocystis				
	Methylosinus				
	Microbacteriaceae	Microbacteriaceae			
	Microbacterium				
	Microbulbifer				
	Micrococcus				
	Micromonospora				
	Mycobacterium	Mycobacterium			
	Nesterenkonia				
	Nitratireductor	Nitratireductor			
	Nitrosococcus	Nitrosococcus			
	Novosphingobium	Novosphingobium			
	Oceanibaculum				
	Oceanicaulis	Oceanicaulis			
	Oceanicola	Oceanicola	Oceanicola	Oceanicola	
	Oceaniovalibus	Oceaniovalibus	Oceaniovalibus		
	Octadecabacter	Octadecabacter			
	Oscillatoria				
	Paenibacillus				
	Pantoea	Pantoea			

	Paracoccus	Paracoccus	Paracoccus		
	Patulibacter	Phaeobacter			
	Pelagibacterium				
	Phaeobacter		Phaeobacter	Phaeobacter	Phaeobacter
	Phaeodactylum				
	Phenylobacterium	Phenylobacterium			
	Polaribacter	Polaribacter			
	Polaromonas	Polaromonas			
	Polynucleobacter	Polynucleobacter			
	Ponticaulis				
	Porphyrobacter	Porphyrobacter			
	Prochlorococcus	Prochlorococcus	Prochlorococcus	Prochlorococcus	Prochlorococcus
	Propionibacterium				
	Pseudoalteromonas	Pseudoalteromonas	Pseudoalteromonas	Pseudoalteromonas	Pseudoalteromonas
	Pseudomonas	Pseudomonas	Pseudomonas	Pseudomonas	Pseudomonas
	Pseudovibrio	Pseudovibrio			
	Psychroflexus	Psychroflexus			
	Ralstonia				
	Ramlibacter	Ramlibacter	Ramlibacter	Ramlibacter	
	Renibacterium	Renibacterium			
	Rhizobium	Rhizobium			
	Rhodanobacter				
	Rhodobacter	Rhodobacter	Rhodobacter	Rhodobacter	Rhodobacter
	Rhodobacteraceae	Rhodobacteraceae	Rhodobacteraceae	Rhodobacteraceae	Rhodobacteraceae
	Rhodococcus	Rhodococcus			
	Rhodopirellula	Rhodopirellula	Rhodopirellula	Rhodopirellula	Rhodopirellula
	Robiginitomaculum	Robiginitomaculum			
	Roseobacter	Roseobacter	Roseobacter	Roseobacter	Roseobacter
	Roseovarius	Roseovarius	Roseovarius	Roseovarius	
	Ruegeria	Ruegeria	Ruegeria	Ruegeria	Ruegeria
	Saccharina	Saccharina	Saccharina		

	Salinisphaera				
	Sandarakinorhabdus				
	Schlesneria	Schlesneria			
	Segetibacter	Segetibacter			
	Serratia	Serratia	Serratia		
	Shinella	Shinella			
	Simiduia	Simiduia			
	Sinorhizobium				
	Sphingobium	Sphingobium	Sphingobium	Sphingobium	Sphingobium
	Sphingomonadaceae				
	Sphingomonadales	Sphingomonadales			
	Sphingomonas	Sphingomonas	Sphingomonas	Sphingomonas	Sphingomonas
	Sphingopyxis	Sphingopyxis	Sphingopyxis	Sphingopyxis	
	Spirochaeta	Spirochaeta	Spirochaeta		
	Spongiibacter	Spongiibacter			
	Stenotrophomonas	Stenotrophomonas			
	Streptococcus	Streptococcus			
	Streptomyces				
	Sulfitobacter	Sulfitobacter			
	Synechococcus	Synechococcus	Synechococcus	Synechococcus	Synechococcus
	Thiothrix	Thiothrix	Thiothrix	Thiothrix	Thiothrix
	Tistrella	Tistrella			
	Treponema	Treponema			
	Tsukamurella				
	Verrucomicrobia	Verrucomicrobia	Verrucomicrobia	Verrucomicrobia	Verrucomicrobia
	Vibrio	Vibrio	Vibrio	Vibrio	Vibrio
	Xanthobacteraceae	Xanthobacteraceae			
	Xanthomonas				
	Xenorhabdus				

Supplementary Table 6: Prokaryotic community composition across stations. Percentage of T1 average of the three replicates. Values are shown only if >0.5%.

L1	L2	L3	S1	S2	S3	S4	S5	S6	Average
Archaea	**Euryarchaeota**	Thermoplasmata						1.71%	
	Thaumarchaeota	Marine Group I		0.99%					
	Total Archea			1.45%				1.89%	0.65%
Bacteria	**Cyanobacteria**	Cyanobacteria	40.26%	58.79%			12.79%	15.73%	
		Total Cyanobacteria	40.32%	58.86%			12.81%	15.77%	21.34%
	Proteobacteria	Alphaproteobacteria	20.22%	11.28%	20.94%	36.36%	25.82%	22.69%	
		Betaproteobacteria			2.02%	2.28%	1.91%	1.80%	
		Deltaproteobacteria	4.11%	4.11%			0.99%	1.04%	
		Elev-16S-509	0.89%	0.69%			0.67%	0.83%	
		Gammaroteobacteria	25.50%	8.04%	32.32%	26%	29.99%	32.81%	
		Total Proteobacteria	32.95%	24.55%	55.67%	64.91%	59.40%	59.79%	49.55%
	Bacteroidetes	Flavobacteriia	3.21%	3.72%	19.56%	23.86%	11.10%	12.89%	
		Total Bacteroidetes	3.36%	4.17%	20.01%	24.37%	11.44%	13.17%	12.76%
	Actinobacteria	Acidomicrobiia	1.72%	4.52%			0.57%	1.99%	
		Total Actinobacteria	1.73%	4.54%			0.60%	2.05%	1.52%
	Chloroflexi			0.73%				0.59%	
	Deferribacteres			1.36%			0.98%	1.94%	
	Planctomycetes			0.67%				0.55%	
	Verrucomicrobia	Opitutae			0.95%	0.87%	2.44%	3.07%	
		Total Verrucomicrobia	0.50%	0.58%	0.96%	0.88%	2.87%	3.39%	1.53%
	Total Bacteria		79.54%	96.23%	77.40%	90.67%	88.75%	97.45%	88.34%
No blast hit			20.34%	2.31%	22.59%	9.32%	10.82%	0.66%	11.01%

Supplementary Table 7: Eukaryotes community composition across stations. Percentage of T1 average of the three replicates. Values are shown only if >0.5%. Avg: average.

L1	L2	L3	L4	L5	S1	S2	S3	S4	S5	S6	Avg
Eukaryota	**Archaeplastida**	Chloroplastida			2.09%	2.33%	0.70%		3.40%	1.83%	
		Rhodophyceae			0.58%						
		Total Archeaplastida			2.68%	2.42%	0.73%		3.50%	1.94%	2.07
	Cryptophyceae	Cryptomonadales			0.67%	1.32%	2.15%	2.02%			
		Total Cryptophyceae			0.72%	1.47%	2.21%	2.02%	0.57%		1.32
	Haptophyta	Pavlovophyceae			1.09%						
		Prymnesiophycea			0.87%	1.06%	18.82	27.10	3.60%	3.18%	
		Prymnesiophycea		*Phaeocystis*	0.05		17.71	26.45	0.96	0.62	
		Total Haptophyta			2.33%	1.15%	18.86	27.17	3.78%	3.21%	9.42
	Opisthokonta	Holozoa					0.52%	0.51%			
	SAR	Alveolata	Dinoflagellata	Dinophyceae	33.54%	31.61	27.51	30.40	31.09	17.39	
				uncultured	4.47%	8.79%	9.16%	5.75%	10.90	6.65%	
				Total	41.35%	42.14	37.57	36.47	43.78	22.58	37.32
			Protoalveolat a	Syndiniales	42.37%	40.44	33.66	22.55	35.13	57.50	
				Perkinsidae	1.43%				0.56%		
				Total	43.86%	40.73	33.8%	22.57	35.68	57.73	39.06
			Total Alveolata		91.66%	88.14	74.79	68.98	86.63	90.93	83.52
		Rhizaria			1.14%	2.84%	1.34%		2.30%	2.29%	
		Stramenopiles						0.60%			
		Total SAR			92.92%	91.12	76.61	69.99	89.19	93.27	85.52
	uncultured marine				0.77%	1.79%	0.62%		2.08%	0.51%	
	Total Eukaryota				99.72%	99.09	99.77	99.99	99.68	99.84	99.68
No blast hit						0.91%					

Supplementary Table 8: Percentage of contigs annotated using the virus db. Values are shown only if > 0.5%.

Order	**Family**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**Average**
Caudovirales	*Myoviridae*	26.03%	31.53%	22.70%	19.73%	23.86%	20.56%	
	Podoviridae	10.33%	21.14%	10.27%	11.89%	20.01%	9.90%	
	Siphoviridae	22.35%	16.16%	21.72%	24.32%	18.93%	24.87%	
	Unassigned	0.71%	2.31%	1.01%	0.77%	1.02%	1.29%	
	Total *Caudovirales*	59.41%	71.14%	55.71%	56.72%	63.82%	56.62%	60.57%
NCLDV	*Ascoviridae*	1.70%	1.28%	1.72%	1.60%	1.35%		
	Iridoviridae	0.57%					0.54%	
	Marseilleviridae	0.57%	0.57%			0.61%	0.65%	
	Mimiviridae	7.92%	3.64%	9.56%	9.56%	5.44%	7.00%	
	Pandoraviridae	2.69%	1.52%	2.77%	3.04%	2.17%	3.23%	
	Phycodnaviridae	13.01%	7.52%	16.28%	16.42%	10.14%	15.39%	
	Poxviridae	1.56%	0.59%	1.59%	0.77%	0.95%	1.51%	
	Total *NCLDV*	28.15%	15.32%	32.67%	32.17%	20.92%	28.85%	26.35%
	Herpesviridae		3.12%			4.55%	2.37%	
	Total *Herpesvirales*		3.42%	0.51%		4.78%	2.58%	1.97%
Other	*Baculoviridae*				0.83%	0.66%	0.54%	
	Chlorovirus	1.70%	0.66%	1.35%	1.38%	0.66%	0.65%	
	Inoviridae						0.75%	
	Unassigned	9.62%	8.51%	8.28%	7.46%	8.28%	9.36%	
	Total *Other*	12.31%	10.05%	10.95%	10.45%	10.32%	11.73%	10.97%

Appendix I: Primer and DNA concentration

Station	Replicate	Forward primer	Reverse primer	DNA amount ng/µl
S1a 16S	Rep1	515F	806R7	27.1
	Rep1	515F	806R7	21.64
	Rep1	515F	806R7	32.51
	Rep2	515F	806R10	10.13
	Rep2	515F	806R10	10.25
	Rep2	515F	806R10	11.44
	Rep3	515F	806R15	1.82
	Rep3	515F	806R15	1.94
	Rep3	515F	806R15	1.83
S1a 18S	Rep1	1391F	EukB6	6.41
	Rep1	1391F	EukB6	7.97
	Rep1	1391F	EukB6	4.49
	Rep2	1391F	EukB16	2
	Rep2	1391F	EukB16	1.68
	Rep2	1391F	EukB16	1.82
	Rep3	1391F	EukB23	2.07
	Rep3	1391F	EukB23	1.59
	Rep3	1391F	EukB23	1.47
S1a permeate				18.5
S1b 16S	Rep1	515F	806R1	2.21
	Rep1	515F	806R1	1.54
	Rep1	515F	806R1	1.91
	Rep2	515F	806R2	3.85
	Rep2	515F	806R2	3.76
	Rep2	515F	806R2	4.60
	Rep3	515F	806R7	2.19
	Rep3	515F	806R7	1.70
	Rep3	515F	806R7	1.87
S2 16S	Rep1	515F	806R4	27.91
	Rep1	515F	806R4	38.52
	Rep1	515F	806R4	26.42
	Rep2	515F	806R13	14.33
	Rep2	515F	806R13	10.3
	Rep2	515F	806R13	5.34
	Rep3	515F	806R20	20.02
	Rep3	515F	806R20	16.09
	Rep3	515F	806R20	23.59
S2 18S	Rep1	1391F	EukB2	4.07
	Rep1	1391F	EukB2	3.89
	Rep1	1391F	EukB2	6.09
	Rep2	1391F	EukB7	2.28

	Rep2	1391F	EukB7	5.95
	Rep2	1391F	EukB7	4.03
	Rep3	1391F	EukB21	5.28
	Rep3	1391F	EukB21	7.05
	Rep3	1391F	EukB21	5.77
S2 permeate				14.9
S3 16S	Rep1	515F	806R6	4.4
	Rep1	515F	806R6	2.69
	Rep1	515F	806R6	3.87
	Rep2	515F	806R18	3.77
	Rep2	515F	806R18	4.46
	Rep2	515F	806R18	6.17
	Rep3	515F	806R24	10.11
	Rep3	515F	806R24	7.03
	Rep3	515F	806R24	8.13
S3 18S	Rep1	1391F	EukB10	5.82
	Rep1	1391F	EukB10	4.73
	Rep1	1391F	EukB10	3.89
	Rep2	1391F	EukB14	1.8
	Rep2	1391F	EukB14	2.94
	Rep2	1391F	EukB14	2.09
	Rep3	1391F	EukB19	7.75
	Rep3	1391F	EukB19	5.5
	Rep3	1391F	EukB19	7.22
S3 permeate				17.1
S4 16S	Rep1	515F	806R5	10.04
	Rep1	515F	806R5	11.08
	Rep1	515F	806R5	9.29
	Rep2	515F	806R11	5.64
	Rep2	515F	806R11	5.54
	Rep2	515F	806R11	6.39
	Rep3	515F	806R22	5.81
	Rep3	515F	806R22	5.2
	Rep3	515F	806R22	3.56
S4 18S	Rep1	1391F	EukB3	13.99
	Rep1	1391F	EukB3	10.08
	Rep1	1391F	EukB3	12
	Rep2	1391F	EukB13	13.01
	Rep2	1391F	EukB13	8.8
	Rep2	1391F	EukB13	8.11
	Rep3	1391F	EukB20	13.31
	Rep3	1391F	EukB20	11.64
	Rep3	1391F	EukB20	12.27
S4 permeate				16
S5 16S	Rep1	515F	806R9	5.11
	Rep1	515F	806R9	5.03
	Rep1	515F	806R9	3.53
	Rep2	515F	806R12	5.11

	Rep2	515F	806R12	4.09
	Rep2	515F	806R12	3.49
	Rep3	515F	806R21	4.15
	Rep3	515F	806R21	4.62
	Rep3	515F	806R21	2.97
S5 18S	Rep1	1391F	EukB11	11.09
	Rep1	1391F	EukB11	10.84
	Rep1	1391F	EukB11	9.07
	Rep2	1391F	EukB15	4.56
	Rep2	1391F	EukB15	5.72
	Rep2	1391F	EukB15	4.71
	Rep3	1391F	EukB24	7.25
	Rep3	1391F	EukB24	11.43
	Rep3	1391F	EukB24	9.82
S5 permeate				17.7
S6 16S	Rep1	515F	806R3	8.27
	Rep1	515F	806R3	5.86
	Rep1	515F	806R3	4.64
	Rep2	515F	806R8	5.44
	Rep2	515F	806R8	6.24
	Rep2	515F	806R8	5.04
	Rep3	515F	806R19	9.89
	Rep3	515F	806R19	6.91
	Rep3	515F	806R19	8.87
S6 18S	Rep1	1391F	EukB4	21.85
	Rep1	1391F	EukB4	14.37
	Rep1	1391F	EukB4	19.78
	Rep2	1391F	EukB17	9.57
	Rep2	1391F	EukB17	11.35
	Rep2	1391F	EukB17	11.06
	Rep3	1391F	EukB22	6.2
	Rep3	1391F	EukB22	5.34
	Rep3	1391F	EukB22	6.28
S6 permeate				24.7
S7 16S	Rep1	515F	806R3	1.064
	Rep1	515F	806R3	3.922
	Rep1	515F	806R3	1.151
	Rep2	515F	806R5	0.587
	Rep2	515F	806R5	0.391
	Rep2	515F	806R5	1.936
	Rep3	515F	806R8	2.434
	Rep3	515F	806R8	3.899
	Rep3	515F	806R8	1.141
S8 16S	Rep1	515F	806R9	1.663
	Rep1	515F	806R9	1.59
	Rep1	515F	806R9	1.801
	Rep2	515F	806R13	1.911
	Rep2	515F	806R13	1.015

	Rep2	515F	806R13	0.598
	Rep3	515F	806R21	2.009
	Rep3	515F	806R21	4.432
	Rep3	515F	806R21	3.204
S9 16S	Rep1	515F	806R1	0.93
	Rep1	515F	806R1	1.27
	Rep1	515F	806R1	1.29
	Rep2	515F	806R16	1.61
	Rep2	515F	806R16	0.97
	Rep2	515F	806R16	1.06
	Rep3	515F	806R23	1.12
	Rep3	515F	806R23	1.38
	Rep3	515F	806R23	0.98

Appendix II: R scripts

```
################################
##### OTU analysis R script#####
####     16s and 18s        ####
################################

# Set and check working directory CHANGE TO WORKING DIR WHICH CONTAIN SAVED
OTUTABLE FILE IN TXT FORMAT (TAB_DELIMITED)
setwd(dir = "W:/")

#Control working directory
getwd()

#Clear workspace
rm(list=ls())

# Close all graphics windows
graphics.off()

### Load existing packages and install packages if non-existing
ipak <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)
}
packages <-
c("sfsmisc","vegan","clusterSim","cluster","RColorBrewer","ggplot2","gridExtra","am
ap","plyr","picante","car","lattice","scales","GGally","gclus","reshape","reshape2"
, "gtable", "gridExtra", "Rmisc")

ipak(packages)

#Make sure samples are columns and OTUs rows
#repeats all these steps for the 16S dataset

readfile_xxS = "OTU_Contingency_Table_L12.txt"

###Read OTU count data
tab <- read.delim(readfile_xxS, row.names = 1)
id <- rownames(tab)
reads <- as.matrix(tab[,1:(ncol(tab)-1)])
taxonomy <- as.matrix(tab[,ncol(tab) - 0])
size <- apply(reads, 2, sum) # number of reads per sample
sample <- colnames(reads)

dim(tab)
str(tab)

#number of individual (sequences), Average reads and standard deviation
size
sum(size)
mean(size)
sd(size)
min(size)
max(size)

write.table(size,"read_distribution.txt",sep="\t")

norm_reads <- reads
for (i in 1:ncol(reads)) {
  norm_reads[,i] <- reads[,i]/sum(reads[,i])
}
```

```r
#make subsampled matrix, n=subsampling level
#for n use the minimal value calculated in the step above
#you will now get the same number of sequences for each sample
n=min(size)
#view n and check its value
n
subs_reads = matrix(ncol = ncol(reads), nrow = nrow(reads))
subs_reads = t(rrarefy(t(reads), n))

#Check if subsampling has been performed correctly
#they should all have the same number as n
colSums(subs_reads)

#Transpose needed for diversity analyses
norm_reads<-t(norm_reads)
subs_reads <-t(subs_reads)


### ALPHA DIVERSITY ###
#index Shannon, only useful if data has been subsetted!
Shannon_index<-diversity(subs_reads, index="shannon")
write.table(Shannon_index,"Alpha_Shannon_index.txt",sep="\t")
pdf("Alpha_Shannon_diversity.pdf")
par(mfrow=c(1,1))

barplot(Shannon_index,names.arg=rownames(subs_reads),las=2,
        col = c("darksalmon",  "brown4", "darkslategray1",  "dodgerblue4",
"greenyellow", "green4" ))
title(main = "Alpha_Shannon_index", font = 4)
dev.off()

#Evenness
evenness<-diversity(subs_reads, index="shannon")/log(specnumber(subs_reads))
pdf("Alpha_Shannon_diversity_evenness.pdf")
#color plot
barplot(evenness,names.arg=rownames(subs_reads),las=2,
        col = c("darksalmon",  "brown4", "darkslategray1",  "dodgerblue4",
"greenyellow", "green4" ))
title(main = "Alpha_Shannon_diversity_evenness", font = 4)
dev.off()

write.table(evenness,"Alpha_Evenness.txt",sep="\t")

#Observed richness, Chao1 and ACE
richness<-estimateR(subs_reads)
write.table(richness,"Alpha_Richness.txt",sep="\t")
richness<-t(richness)
richness<-data.frame(richness)
attach(richness)

pdf("Alpha_Richness.pdf")
barplot(S.obs,names.arg=rownames(subs_reads),las=2,
        col = c("darksalmon",  "brown4", "darkslategray1",  "dodgerblue4",
"greenyellow", "green4" ),
        main="S.obs")
barplot(S.chao1,names.arg=rownames(subs_reads),las=2,
        col = c("darksalmon",  "brown4", "darkslategray1",  "dodgerblue4",
"greenyellow", "green4" ),
        main="S.chao1")
barplot(S.ACE,names.arg=rownames(subs_reads),las=2,
        col = c("darksalmon",  "brown4", "darkslategray1",  "dodgerblue4",
"greenyellow", "green4" ),
        main="S.ACE")

dev.off()


#Rank abundance analys
```

```
#All models (separate sites)
radfit_reads<-radfit(subs_reads)
plot(radfit_reads, legend=T)
pdf("Alpha_radfit_smallersizefile.pdf", pointsize = 10)
plot(radfit_reads, legend=T)
dev.off()

#Lognormal model (all sites)
rad.lognormal(subs_reads)
plot(rad.lognormal(subs_reads))
pdf("Alpha_lognormal.pdf")
plot(rad.lognormal(subs_reads))
dev.off()


### BETA DIVERSITY ###
#Distance matrix - Bray-curtis
distmatris<- vegdist(norm_reads,method="bray")
capture.output(distmatris, file = "Beta_Dist_matris.txt")

#Hierarchical cluster analys (dendrogram)
ag<-agnes(distmatris)
dgr<-as.dendrogram(as.hclust(ag))
plot(dgr, edgePar = list(lwd=2))
pdf("Beta_dendrogram.pdf")
plot(dgr, edgePar = list(lwd=2))
dev.off()

#NMDS (Non-Metric Multidimensional Scaling)
nmdsmatris<-metaMDS(norm_reads)
ordiplot(nmdsmatris, type="points",display="sites", xlim =c(-2,2), ylim = c(-2,2))
abline(h = 0, v = 0)
points(nmdsmatris, "sites", pch=20,
       col = c("darksalmon",  "brown4", "darkslategray1",  "dodgerblue4",
"greenyellow", "green4" ),
       bg="white",cex=4, choices = c(1,2))
text(nmdsmatris, "sites", col="black",cex=0.7)
title(main = "Beta_nMDS", font = 4)

#Make sure to set the xlim and ylim parameters and the save the plot
pdf("Beta_nMDS.pdf")
ordiplot(nmdsmatris, type="points",display="sites", xlim =c(-2,2), ylim = c(-2,2))
abline(h = 0, v = 0)
points(nmdsmatris, "sites", pch=20,
       col = c("darksalmon",  "brown4", "darkslategray1",  "dodgerblue4",
"greenyellow", "green4" ),
       bg="white",cex=4, choices = c(1,2))
text(nmdsmatris, "sites", col="black",cex=0.5)
title(main = "Beta_nMDS", font = 4)
dev.off()


### POPULATIONS  18S ####
reads<-t(reads)

Amoebozoa<-grep("Eukaryota;Amoebozoa", taxonomy)
Charophyta<-grep("Chloroplastida;Charophyta", taxonomy)
Chlorophyta<-grep("Chloroplastida;Chlorophyta", taxonomy)
Rhodophyceae<-grep("Archaeplastida;Rhodophyceae", taxonomy)
Centrohelida<-grep("Eukaryota;Centrohelida", taxonomy)
Cryptophyceae<-grep("Eukaryota;Cryptophyceae", taxonomy)
Excavata<-grep("Eukaryota;Excavata", taxonomy)
Haptophyta<-grep("Eukaryota;Haptophyta", taxonomy)
Opisthokonta<-grep("Eukaryota;Opisthokonta", taxonomy)
Fungi<-grep("Nucletmycea;Fungi", taxonomy)
Ascomycota<-grep("Dikarya;Ascomycota",taxonomy)
Basidiomycota<-grep("Dikarya;Basidiomycota", taxonomy)
Apicomplexa<-grep("Alveolata;Apicomplexa", taxonomy)
```

```
Ciliophora<-grep("Alveolata;Ciliophora", taxonomy)
Dinoflagellata<-grep("Alveolata;Dinoflagellata", taxonomy)
Protalveolata<-grep("Alveolata;Protalveolata", taxonomy)
Rhizaria<-grep("SAR;Rhizaria",taxonomy)
Stramenopiles<-grep("SAR;Stramenopiles",taxonomy)
NoBlast<-grep("No blast hit", taxonomy)


sum_reads_Amoebozoa<-rowSums(reads[,Amoebozoa])/(size)
sum_reads_Charophyta<-rowSums(reads[,Charophyta])/(size)
sum_reads_Chlorophyta<-rowSums(reads[,Chlorophyta])/(size)
sum_reads_Rhodophyceae<-rowSums(reads[,Rhodophyceae])/(size)
sum_reads_Centrohelida<-rowSums(reads[,Centrohelida])/(size)
sum_reads_Cryptophyceae<-rowSums(reads[,Cryptophyceae])/(size)
sum_reads_Excavata<-rowSums(reads[,Excavata])/(size)
sum_reads_Haptophyta<-rowSums(reads[,Haptophyta])/(size)
sum_reads_OpisthokontaNofungi<-(rowSums(reads[,Opisthokonta])-
rowSums(reads[,Fungi]))/(size)
sum_reads_Ascomycota<-rowSums(reads[,Ascomycota])/(size)
sum_reads_Basidiomycota<-rowSums(reads[,Basidiomycota])/(size)
sum_reads_OtherFungi<-(rowSums(reads[,Fungi])-(rowSums(reads[,Ascomycota]))-
(rowSums(reads[,Basidiomycota])))/(size)
sum_reads_Apicomplexa<-rowSums(reads[,Apicomplexa])/(size)
sum_reads_Ciliophora<-rowSums(reads[,Ciliophora])/(size)
sum_reads_Dinoflagellata<-rowSums(reads[,Dinoflagellata])/(size)
sum_reads_Proalveolata<-rowSums(reads[,Protalveolata])/(size)
sum_reads_Rhizaria<-rowSums(reads[,Rhizaria])/(size)
sum_reads_Strametopiles<-rowSums(reads[,Stramenopiles])/(size)
sum_reads_NoBlast<-rowSums(reads[,NoBlast])/(size)

#Grand sum of these groups - to find out what all other groups correspond to
grand_sum<-
colSums(t(sum_reads_Amoebozoa+sum_reads_Charophyta+sum_reads_Chlorophyta+sum_reads_
Rhodophyceae+sum_reads_Centrohelida+sum_reads_Cryptophyceae+sum_reads_Excavata+sum_
reads_Haptophyta+sum_reads_OpisthokontaNofungi+sum_reads_Ascomycota+sum_reads_Basid
iomycota+sum_reads_OtherFungi+sum_reads_Apicomplexa+sum_reads_Ciliophora+sum_reads_
Dinoflagellata+sum_reads_Proalveolata+sum_reads_Rhizaria+sum_reads_Strametopiles+su
m_reads_NoBlast))*size

sum_reads_Others<-(size-grand_sum)/(size)

#Create dataframe for the groups
Taxon<-
t(rbind(sum_reads_Amoebozoa,sum_reads_Charophyta,sum_reads_Chlorophyta,sum_reads_Rh
odophyceae,sum_reads_Centrohelida,sum_reads_Cryptophyceae,sum_reads_Excavata,sum_re
ads_Haptophyta,sum_reads_OpisthokontaNofungi,sum_reads_Ascomycota,sum_reads_Basidio
mycota,sum_reads_OtherFungi,sum_reads_Apicomplexa,sum_reads_Ciliophora,sum_reads_Di
noflagellata,sum_reads_Proalveolata,sum_reads_Rhizaria,sum_reads_Strametopiles,sum_
reads_Others,sum_reads_NoBlast))


#########################
###POPULATION 16S ###
#chloroplasts and mithocondria sequences have been remove prior to this step
Cyanobacteria<-grep("Cyanobacteria; Cyanobacteria", taxonomy)
Bacteroidetes<-grep("Bacteria; Bacteroidetes", taxonomy)
Actinobacteria<-grep("Bacteria; Actinobacteria", taxonomy)
Verrucomicrobia<-grep("Bacteria; Verrucomicrobia", taxonomy)
Alphaproteobacteria<-grep("Bacteria; Proteobacteria; Alphaproteobacteria",
taxonomy)
Betaproteobacteria<-grep("Bacteria; Proteobacteria; Betaproteobacteria", taxonomy)
Gammaproteobacteria<-grep("Bacteria; Proteobacteria; Gammaproteobacteria",
taxonomy)
Deltaproteobacteria<-grep("Bacteria; Proteobacteria; Deltaproteobacteria",
taxonomy)
Planctomycetes<-grep("Bacteria; Planctomycetes", taxonomy)
Epsilonproteobacteria<-grep("Bacteria; Proteobacteria; Epsilonproteobacteria",
taxonomy)
```

```
NoBlast<-grep("No blast hit", taxonomy)
Archaea<-grep("Archaea", taxonomy)

#Sum up reads from each group relative of each other
#calculate relative aboundance of each group
sum_reads_Cyanobacteria<-rowSums(reads[,Cyanobacteria])/(size)
sum_reads_Bacteroidetes<-rowSums(reads[,Bacteroidetes])/(size)
sum_reads_Actinobacteria<-rowSums(reads[,Actinobacteria])/(size)
sum_reads_Verrucomicrobia<-rowSums(reads[,Verrucomicrobia])/(size)
sum_reads_Alphaproteobacteria<-(rowSums(reads[,Alphaproteobacteria]))/(size)
sum_reads_Betaproteobacteria<-rowSums(reads[,Betaproteobacteria])/(size)
sum_reads_Gammaproteobacteria<-rowSums(reads[,Gammaproteobacteria])/(size)
sum_reads_Deltaproteobacteria<-rowSums(reads[,Deltaproteobacteria])/(size)
sum_reads_Planctomycetes<-rowSums(reads[,Planctomycetes])/(size)
sum_reads_Epsilonproteobacteria<-rowSums(reads[,Epsilonproteobacteria])/(size)
sum_reads_NoBlast<-rowSums(reads[,NoBlast])/(size)
sum_reads_Archaea<-rowSums(reads[,Archaea])/(size)

#Grand sum of these groups - to find out what all other groups correspond to
#how many orders you have to calculate % for the barplot
#NOTA BENE: if you modify the list above (i.e. add remove something) you'll need to
change it here as well

grand_sum<-
colSums(t(sum_reads_Cyanobacteria+sum_reads_Bacteroidetes+sum_reads_Actinobacteria+
sum_reads_Verrucomicrobia+sum_reads_Alphaproteobacteria+sum_reads_Betaproteobacteri
a+sum_reads_Gammaproteobacteria+sum_reads_Deltaproteobacteria+sum_reads_Planctomyce
tes+sum_reads_Epsilonproteobacteria+sum_reads_NoBlast+sum_reads_Archaea))*size

sum_reads_Others<-(size-grand_sum)/(size)

#Create dataframe for the groups
Taxon<-
t(rbind(sum_reads_Cyanobacteria,sum_reads_Bacteroidetes,sum_reads_Actinobacteria,su
m_reads_Verrucomicrobia,sum_reads_Alphaproteobacteria,sum_reads_Betaproteobacteria,
sum_reads_Gammaproteobacteria,sum_reads_Deltaproteobacteria,sum_reads_Planctomycete
s,sum_reads_Others,sum_reads_Epsilonproteobacteria,sum_reads_NoBlast,sum_reads_Arch
aea))

#######################
#Check that the value for rowSums is equal to 100% so all values to sum up to 1
rowSums(Taxon)

#Export Taxon to tab delimited file
write.table(Taxon,"pre_Taxon.txt")
Taxon_readfile<-read.table("pre_Taxon.txt",sep="",header=TRUE)
Taxon<-data.frame(Date=rownames(Taxon_readfile),Taxon_readfile)
taxon_final<-cbind(stack(Taxon),rownames(norm_reads))
colnames(taxon_final) <- c("Rel.Abund","Group","Treatment")

#Export Taxon to tab delimited file
write.table(taxon_final,"Taxonomy.txt",sep="\t")


#run preliminary plot
ggplot(taxon_final, aes(x = Treatment)) + geom_bar(aes(weight=Rel.Abund, fill =
Group)) + scale_fill_manual(values = rev(rainbow(20))) +
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))

# create color paletteup for 25 groups
c25 <- c("dodgerblue2","#E31A1C", # red
         "green4",
         "#6A3D9A", # purple
         "#FF7F00", # orange
         "black","gold1",
         "skyblue2","#FB9A99", # lt pink
         "palegreen2",
         "#CAB2D6", # lt purple
         "#FDBF6F", # lt orange
         "gray70", "khaki2",
         "maroon","orchid1","deeppink1","blue1","steelblue4",
         "darkturquoise","green1","yellow4","yellow3",
         "darkorange4","brown")

#Export Taxon to tab delimited file
write.table(taxon_final,"Taxonomy.txt",sep="\t")

#test the plot with the new colors
ggplot(taxon_final, aes(x = Treatment)) + geom_bar(aes(weight=Rel.Abund, fill =
Group)) + scale_fill_manual(values = rev(c25)) +
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))

pdf("Taxonomy.pdf")
ggplot(taxon_final, aes(x = Treatment)) + geom_bar(aes(weight=Rel.Abund, fill =
Group)) + scale_fill_manual(values = rev(c25)) +
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
dev.off()

#open the file and
#made changes to add locations: A = SWI; B= SO; C= SEI
taxon_final_test <- read.table("Taxonomy_loc.txt",sep="\t",  header = T, as.is=T )

#run preliminary plot
ggplot(taxon_final_test, aes(x = Treatment)) + geom_bar(aes(weight=Rel.Abund, fill
= Group)) + scale_fill_manual(values = rev(rainbow(20))) +
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))

#improve the barplot, separate plot by location

pdf("Taxonomy_locations.pdf")

ggplot(taxon_final_test, aes(x = Treatment)) + geom_bar(aes(weight=Rel.Abund, fill
= Group)) +
  scale_fill_manual(values = rev(c25)) + xlab ("Stations") + ylab("Rel.Abund")+
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5),legend.key.size =
unit(0.5, "cm")) +
  guides(fill=guide_legend(ncol=1)) +
  facet_grid(~Location, scales = "free" )
dev.off()

#Plot OTUs individually
norm_reads<-t(norm_reads)
#Plot OTUs individually
# set samples to include
these <- 1:length(sample)

# set sample names to show on x-axis
xnames <- sample[these]

#select the otus to include the 200 most abundant
selected_otus <- sort(apply(norm_reads[,these], 1, mean), decreasing = TRUE,
index.return = TRUE)$ix[1:200]
selected_reads <- norm_reads[selected_otus,]

ynames<-rownames(selected_reads)
xnames<-colnames(selected_reads)
taxonomy_200<-taxonomy[selected_otus]

#Plot these OTUs
for ( i in seq(1, nrow(selected_reads) )){
  barplot(selected_reads[i,],ylab="relative
abund.",names.arg=sort(xnames),las=2,cex.main=0.6,main=taxonomy_200[i])
  mtext(ynames[i],side=3)
}
pdf("plotall_top_200_OTUs.pdf",pointsize = 10)
par(mfrow=c(2,1))
```

```
for ( i in seq(1, nrow(selected_reads) )){
  barplot(selected_reads[i,],ylab="relative
abund.",names.arg=sort(xnames),las=2,cex.main=0.6,main=taxonomy_200[i])
  mtext(ynames[i],side=3)
}
dev.off()

write.table(selected_reads, "selected_reads.txt",sep="\t")

###PERMANOVA###
## file for permanova
write.table(reads, "Permanova.txt",sep="\t")
write.table(subs_reads, "subs_reads_Permanova.txt",sep="\t")

#prepare the file for Permanova
#Add locations and temperatures (or other values if presents)
OTU <- read.table("Permanova.txt", sep="\t", row.names = 1)

View(OTU)
###add a column
OTU.data <- data.frame(OTU)
#check values
ncol(OTU.data)

OTU.data$Location <- c("SWI", "SWI","SWI","SWI","SWI","SWI",
                       "SO","SO","SO","SO","SO","SO",
                       "SEI","SEI","SEI","SEI","SEI","SEI")
#check values
ncol(OTU.data)

#put the column as first
dataset <- OTU.data[,c(24271, 1: 24270)]
View(dataset)

#add stations and temperature
dataset$Temperature <- c(20.83, 20.83, 20.83, 19.9796, 19.9796, 19.9796,
                         1.377, 1.377, 1.377, 1.236, 1.236, 1.236,
                         16.2288, 16.2288, 16.2288, 12.9472, 12.9472, 12.9472)

#check values
ncol(dataset)

dataset_new <- dataset[,c(1, 24272, 2: 24271)]
ncol(dataset_new)
View(dataset_new)

write.table(dataset_new, "Permanova_adjusted-with-locations-and-
temperature.txt",sep="\t")

#Load the file for Permanova
OTU<-read.table("Permanova_adjusted-with-locations-and-temperature.txt")

View(OTU)
ncol(OTU)
y <- OTU[, 3:24272] #to delimite the OTU data (all the OTUs columns)
View(y)

#Location and temperature as factors (you can run the permutation depending on the
values you want to test)
perm_I<-adonis(y~OTU$Location*Temperature, data=OTU,
permutations=999,method="bray", contr.unordered='contr.sum')
perm_I #view the result of the analyses


####ANOVA#########
#file should look like:
#Sample | Station | Replicate | S.Obs | Shannon | Eveness
#values have been computed previously
```

```
#two-way ANOVA, own
library(car)
d3<-read.csv("for_ANOVA.csv", header=TRUE)
View(d3)
mod.ok <- lm(Sobs ~  Location*Temperature, data=d3)
qqnorm(resid(mod.ok))
plot(fitted(mod.ok),resid(mod.ok))
shapiro.test(resid(mod.ok))

aov_Sobs <- aov(Sobs ~ Location*Temperature, data = d3)
summary(aov_Sobs)
TukeyHSD(aov_Sobs)


#check the effect of station
mod.ok <- lm(Sobs ~  Station, data=d3)
qqnorm(resid(mod.ok))
plot(fitted(mod.ok),resid(mod.ok))
shapiro.test(resid(mod.ok))

aov_Sobs <- aov(Sobs ~ Station, data = d3)
summary(aov_Sobs)
TukeyHSD(aov_Sobs)


#### analyses are done ####
```

```
################################
#####LikelihoodRatio Testing####
################################

#testing for 3 areas you will only need to repeat the test
#area 1 vs area 2
#area 1 vs area 3
#area 2 vs area 3

# read in the data
# A) The area 1 data
# B) The area 2 data


# Set and check working directory FILE IN TXT FORMAT, (TAB_DELIMITED)
setwd(dir = "W:file.txt")


area1<- read.csv(file=file.choose(), header = T, as.is = T, sep = ",")
area2<- read.csv(file=file.choose(), header = T, as.is = T, sep = ",")

all<-merge(area1, area2)
total.order<-apply(as.matrix(all[,2:5]), 1,sum)
total.sample<-apply(as.matrix(all[,2:5]), 2, sum)
total.overall<-sum(total.order)
p.hat.null

output.summary<-function(ct.table){
  total.order<-apply(as.matrix(ct.table[,2:ncol(ct.table)]), 1,sum)
  total.sample<-apply(as.matrix(all[,2:ncol(ct.table)]), 2, sum)
  total.overall<-sum(total.order)
  p.hat.null<- total.order/total.overall
  return(list(total.order, total.sample, total.overall, p.hat = p.hat.null))
}

phat.area1<-output.summary(area1)$p.hat
phat.area2<-output.summary(area2)$p.hat
phat.all<-output.summary(all)$p.hat

get.loglikelihood<-function(x,phat){
  n<-sum(x)
  tempA <- sum(log(c(1:n)))
  tempB <-log(x)
  tempC <-x*log(phat)
  return(tempA - sum(tempB) + sum(tempC))
}

# x,p for each sample under the null.
get.loglikelihood.homogeneous.p <-function(phat,dat){
  loglikelihood <- 0
  for(j in 2:ncol(dat)){
    loglikelihood<- loglikelihood + get.loglikelihood(x = dat[,j], phat = phat)
  }
  return(loglikelihood)
}

null.loglike<-get.loglikelihood.homogeneous.p(phat = phat.all, dat = all)

alternative.loglike<- get.loglikelihood.homogeneous.p(phat = phat.area1, dat =
area1) +
  get.loglikelihood.homogeneous.p(phat = phat.area2, dat = area2)

#df will have to be set depending on your data
lrs<- 2*(alternative.loglike - null.loglike)
pchisq(q = lrs, df = 3, ncp = 0, lower.tail = F, log.p = FALSE)
```

Appendix III: Nextera adapters

Full-length indexed PCR product (green indicates library insert)

```
5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXX-//-XXXXXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG
   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGATGTGCTGCGAGAAGGCTAGAxxxxxx-//-xxxxxxTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNTAGAGCATACGGCAGAAGACGAAC-5’
```

underlining indicates sequences identical to flow cell oligos

Sequencing reads

```
                  Read 1 5’-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-> Index read 5’-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC->
5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXX-//-XXXXXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG
   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGATGTGCTGCGAGAAGGCTAGAxxxxxx-//-xxxxxxTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNTAGAGCATACGGCAGAAGACGAAC-5’
                                                                           <-TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG-5’ Read 2
```

Sequencing Adapter sequences

```
5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
5’-AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG
```

R/C of Sequencing adapter sequences

```
5’- GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
5’- CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
```